# Modelos ocultos de Markov para el reconocimiento automático del habla

Una breve introducción

Sugerencias y correcciones a:
d.milone@ieee.org
1 de marzo de 2004

Los modelos ocultos de Markov (MOM) constituyen una de las técnicas que se ha utilizado con más éxito en el reconocimiento automático del habla (RAH). Principalmente, esta técnica ha permitido modelar adecuadamente la gran variabilidad en el tiempo de la señal de voz. En la terminología del RAH, con MOM suele hacerse referencia no sólo a la técnica de los modelos ocultos de Markov propiamente dicha, sino también a una larga lista de adaptaciones y técnicas asociadas que se fueron incorporando para solucionar el problema de RAH. En este documento se tratarán los conceptos básicos de los MOM y su aplicación al RAH, desde una perspectiva muy simplificada y conceptual[1].

# 1. Modelos para el reconocimiento del habla

Cuando hablamos de RAH pensamos en un sistema automático que intenta transcribir en lenguaje escrito lo que un locutor ha expresado oralmente. Deben distinguirse en primer lugar los sistemas de reconocimiento del habla de los sistemas de *comprensión* del habla. Suele considerarse que la comprensión del habla es un concepto más amplio, que si bien incluye entre otras partes a un sistema de RAH, su objetivo es capturar la semántica del mensaje y no solamente transcribirlo en texto sino entenderlo correctamente. Comenzamos a ver así las potenciales aplicaciones de un sistema de RAH. En toda interfaz entre el hombre y las máquinas resulta de especial interés aprovechar aquel medio de comunicación que entre los hombres más uso ha tenido. Actualmente la mayoría de la gente sigue tecleando unas 60 palabras por minuto (en el mejor de los casos) cuando podría llegar a pronunciar unas 200 en el mismo tiempo. Las aplicaciones del RAH ya son un lugar común —tanto en ciencia como en ficción— por lo que invitamos al lector a consultar la bilbiografía básica que se encuentra al final de este documento.

Si consideramos el proceso de la comunicación oral podríamos pensar que para cada texto el locutor activa un sistema y da como salida una determinada emisión sonora. Para comenzar a entender como se aplican los MOM al RAH imaginemos que para cada una de las posibles emisiones podemos encontrar un modelo capaz de imitar al sistema activado por el locutor. Es decir, un modelo que sea capaz de generar la misma emisión que generó el locutor a partir del texto que había en su mente. De esta forma vamos a suponer que contamos con tantos modelos como posibles emisiones pueda

[1] Para un tratamiento más formal se ha escrito otro documento más detallado "Fundamentos del reconocimiento automático del habla".

hacer el locutor y, para cada modelo un texto asociado. En caso de que conozcamos perfectamente estos modelos, podríamos utilizar el camino inverso para resolver el problema de RAH. Teniendo una determinada emisión del locutor nos preguntaremos: ¿cuál de todos mis modelos generará el sonido más parecido al que generó el locutor? Al encontrar el modelo que genera el sonido más parecido a la emisión del locutor entonces también habremos encontrado el texto, ya que habíamos dicho que todos los modelos estaban asociados a un determinado texto.

Existen dos observaciones de interés en este planteamiento. En primer lugar se debe entender que la solución propuesta es una solución que no parte de la utilización más corriente de los modelos. Generalmente utilizamos un modelo para obtener determinadas salidas a partir de ciertas entradas. Sin embargo, aquí estamos utilizando muchos modelos y una entrada fija asociada a cada uno (el texto). Luego, dada una señal de voz en particular, vemos cuál de todos genera una salida más parecida y damos como resultado la entrada de ese modelo. En segundo lugar, se puede ver claramente que este planteamiento para la solución del problema de RAH no es totalmente aplicable a casos reales, pues sería necesaria una cantidad infinita de modelos. Este problema se resuelve teniendo en cuenta que: 1) no es totalmente necesario abarcar toda la diversidad del habla (ni nosotros mismos podemos hacerlo) y 2) cada modelo no tiene por qué ser totalmente distinto e independiente de los demás.

El segundo punto puede adquirir mayor relevancia si tenemos en cuenta la organización estructural del habla, donde existe una estructura jerárquica en la que pequeños componentes se combinan para formar otros de mayor complejidad (por ejemplo, simplificando mucho la estructura tenemos: fonemas, palabras, frases). Esto quiere decir que sería posible construir una gran cantidad de modelos combinando un número razonable de pequeñas partes. A continuación veremos como modelar estas pequeñas partes por medio de los MOM y cómo generar grandes modelos a partir de ellas. También veremos cómo buscar el modelo cuya salida más se aproxima a la emisión del locutor y cómo encontrar los parámetros que mejor modelan un conjunto de emisiones para diversos locutores.

## 2. Modelos de autómatas finitos

Los autómatas puede ser utilizados para modelar secuencias temporales de variables discretas. Estos modelos poseen un conjunto de estados que representan las diferentes configuraciones internas en que se pueden encon-

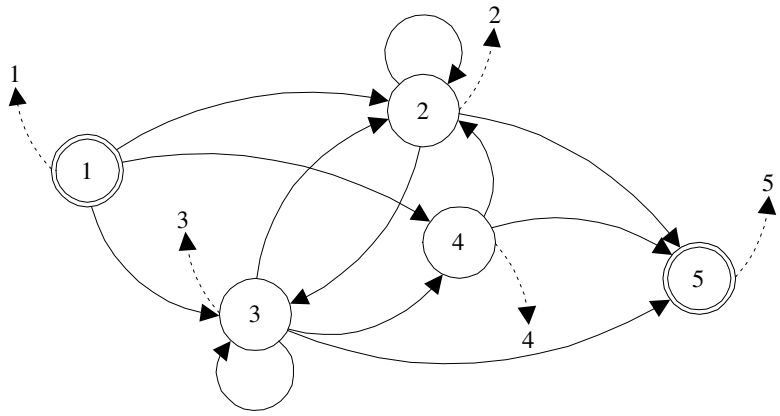


Figura 1. Diagrama de estados para un autómata finito. En este diagrama se puede observar un estado inicial (1), un estado final (5), los estados internos (2..4) y las flechas que indican las posibles transiciones entre los estados. También se han representado las salidas de cada estado en líneas de puntos que, para simplificar, coinciden con el número de estado.

trar. Si el conjunto de estados es finito entonces se habla de de autómatas finitos. Entre los estados debe distinguirse un estado inicial y un estado final. También es necesaria una función de transición de estados que determine la forma en que se realizan los cambios de un estado a otro. Para terminar de ver a los autómatas como un modelo, será necesario especificar entradas y salidas. En estos modelos cada estado puede asociar una salida para la entrada dada. La forma en que se realiza esta asociación da lugar a una gran variedad de autómatas. Por ejemplo, un caso sencillo puede consistir en que cada estado posea una función de salida que selecciona entre los elementos de un conjunto finito de símbolos de salida.

Para representar la estructura interna de un modelo de autómatas suele utilizarse un diagrama de estados como el de la Figura 1. En este diagrama se pueden observar todos los estados, sus salidas y las flechas que indican las posibles transiciones entre ellos.

¿Cómo se puede utilizar este modelo de autómata finito? Para entender un ejemplo sencillo se puede simplificar la función de salida de forma que de como resultado el número del estado y utilizar una función de transición que simplemente elija al estado siguiente como aquel que posee el número más cercano a la entrada actual. Así, dada una secuencia de entrada: 2, 2, 2, 4,

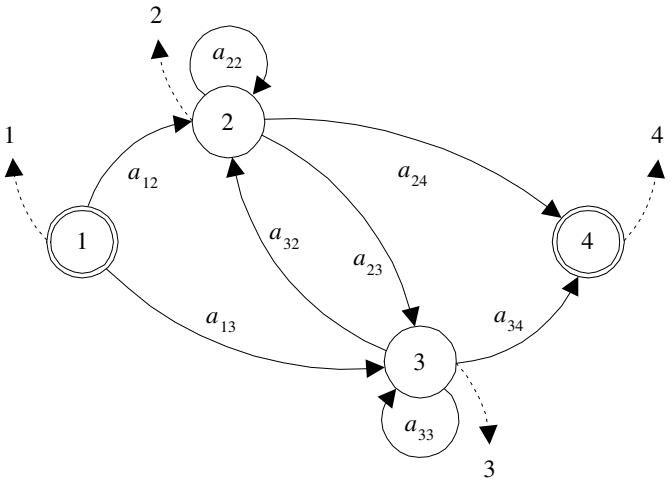


Figura 2. Diagrama de estados para un autómata probabilístico. Las probabilidades de transición desde el estado $i$ al estado $j$ se indican como $a_{ij}$. A cada estado se asocia un símbolo del conjunto finito de salidas. En este ejemplo la salida del estado corresponde simplemente con su número

4, 4 se obtendrá como secuencia de estados: 1, 2, 2, 3, 4, 5 e idénticamente como secuencia de salida: 1, 2, 2, 3, 4, 5.

Otro tipo interesante de autómata es aquél que puede albergar una descripción probabilística del fenómeno que modela. Para estos autómatas es necesario realizar algunas definiciones particulares a partir de los elementos básicos de un autómata finito. En lugar de función de transición de estados se habla de probabilidades de transición entre estados. Es común utilizar para estas probabilidades la notación $a_{ij}$: probabilidad de pasar al estado $j$ dado que se está actualmente en el estado $i$. En cuanto a las salidas de este modelo estadístico, cada estado se asocia a uno de los posibles símbolos de un *conjunto de salidas*. Un ejemplo sencillo se puede observar en la Figura 2.

En este caso también cabe preguntarse: ¿Cómo se pueden utilizar estos modelos de autómatas probabilísticos? Aquí el planteamiento se invierte y se utiliza el modelo de autómatas para encontrar la probabilidad de que una determinada secuencia de salida haya sido generada por él[2]. Es decir, a partir de una secuencia de salidas observadas en el mundo real, se plantea conocer

[2]Estar inversión está orientada hacia la particular forma de utilizar los modelos en RAH, como se discutió en la introducción de esta sección. De esta forma se va introduciendo progresivamente la perspectiva de MOM para RAH.

qué probabilidad existe de que el modelo en cuestión la haya generado. Para dar un ejemplo sencillo se puede suponer que en el modelo de la Figura 2:

$$A = \{a_{ij}\} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 0 & 1/4 & 1/4 & 1/2 \\ 0 & 1/2 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Ahora la pregunta es: ¿Qué probabilidad existe de que este modelo genere la secuencia 1, 2, 2, 3, 2, 4? Para resolver este problema deben considerarse las transiciones de estado $1 \rightarrow 2$, $2 \rightarrow 2$, $2 \rightarrow 3$, $3 \rightarrow 2$ y $2 \rightarrow 4$. Así, se obtiene la probabilidad final para la secuencia mediante la multiplicación:

$$p_{122324} = a_{12}a_{22}a_{23}a_{32}a_{24} = \frac{1}{2}\frac{1}{4}\frac{1}{4}\frac{1}{2}\frac{1}{2} = \frac{1}{128}$$

Este modelo probabilístico es también denominado *modelo de Markov* (MM). Si el tiempo transcurre entre cada transición a intervalos discretos, se dice entonces que se trata de un MM de tiempo discreto. Si además se sigue en la presunción de que las probabilidades de transición sólo dependen de los estados origen y destino, se está en presencia de un proceso de primer orden que suele denominarse cadena de Markov. Como las probabilidades de transición no se modifican con el tiempo también se trata de un sistema invariante en el tiempo o, en la terminología de la teoría de probabilidades, una cadena de Markov homogénea. Finalmente, observando el hecho de que en un MM no se especificaba una entrada, se llega a la denominación de fuente de Markov, muy utilizada en teoría de comunicaciones.

## 3. Modelos ocultos de Markov

En cada estado de un MM se emite un determinado símbolo del conjunto de salidas posibles. Es decir que la función de salida simplemente asigna uno de los símbolos dependiendo del estado en que se encuentre el modelo. Es por esto que un MM es también conocido bajo la denominación de modelo *observable* de Markov: a partir de la salida se puede "observar" en qué estado se encuentra el modelo. El hecho de que en cada estado se pueda observar un único símbolo es una limitación importante que reduce las posibilidades de aplicación de los MM. Para aumentar su capacidad de modelado, se ha propuesto una extensión en donde la función que asocia a

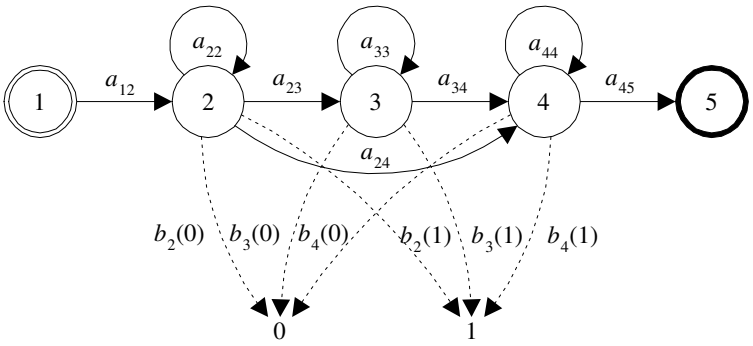


Figura 3. Diagrama de estados para un modelo oculto de Markov comúnmente utilizado en RAH. El estado 1 (estado inicial) y el 5 (estado final) se denominan no emisores. Las flechas en líneas continuas indican las posibles transiciones entre estados. Las flechas en líneas de puntos indican las probabilidades de observación para cada estado. En esta configuración se puede observar la particularidad de que las transiciones se dan solamente de izquierda a derecha.

cada estado una salida sea una distribución de probabilidades sobre todas las posibles salidas. Ahora existirá un nuevo parámetro $b_j(k)$ que describe la probabilidad de que el estado $j$ observe el símbolo $k$ del conjunto de salidas[3]. En estas condiciones nunca se podrá saber con certeza en qué estado esta el modelo observando solamente su salida. El funcionamiento interno del modelo queda "oculto" y es por eso que se lo denomina modelo *oculto* de Markov. Los MOM más utilizados en RAH poseen una estructura muy simple denominada de izquierda a derecha. Un ejemplo de estas estructuras se muestra en la Figura 3.

Si para el modelo de la Figura 3 se dan los parámetros:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 1/3 & 1/5 & 2/3 & 0 \\ 0 & 2/3 & 4/5 & 1/3 & 0 \end{bmatrix}$$

una de las preguntas más importantes está relacionada nuevamente con la probabilidad de generar una secuencia observada: ¿qué probabilidad existe

[3]En algunos casos suele hablarse de probabilidades de *emisión* en lugar de probabilidades de *observación*.

de que este modelo genere la secuencia 0, 0, 1, 0? La respuesta no es tan obvia como en los casos anteriores. En este caso no se puede inferir directamente la secuencia de estados que debería haber seguido el modelo para generar esa salida ya que el modelo está "oculto". Si se analiza un poco más el problema se puede deducir que la secuencia de estados que genera esa secuencia de salida no es única: ahora cada estado puede emitir cualquiera de los símbolos del conjunto de salidas (aunque con distinta probabilidad). Para resolver este problema es necesario analizar todas las posibles secuencias que pasen por 4 estados emisores y sus probabilidades asociadas (véase la Tabla 1). Una forma alternativa para representar estas transiciones de estados es la que se muestra en el diagrama de la Figura 4.

| Secuencias de de estados | Probabilidades de transición | Probabilidades de observación | Probabilidades de la secuencia |
|---|---|---|---|
| 1, 2, 2, 2, 4, 5 | $1\ \frac{1}{4}\frac{1}{4}\frac{1}{2}\frac{1}{2} = \frac{1}{64}$ | $\frac{1}{3}\frac{1}{3}\frac{2}{3}\frac{2}{3} = \frac{4}{81}$ | $\frac{1}{1296}$ |
| 1, 2, 2, 3, 4, 5 | $1\ \frac{1}{4}\frac{1}{4}\frac{1}{2}\frac{1}{2} = \frac{1}{64}$ | $\frac{1}{3}\frac{1}{3}\frac{4}{5}\frac{2}{3} = \frac{8}{135}$ | $\frac{1}{1080}$ |
| 1, 2, 2, 4, 4, 5 | $1\ \frac{1}{4}\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{32}$ | $\frac{1}{3}\frac{1}{3}\frac{1}{3}\frac{2}{3} = \frac{2}{81}$ | $\frac{1}{1296}$ |
| 1, 2, 3, 3, 4, 5 | $1\ \frac{1}{4}\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{32}$ | $\frac{1}{3}\frac{1}{5}\frac{4}{5}\frac{2}{3} = \frac{8}{225}$ | $\frac{1}{900}$ |
| 1, 2, 3, 4, 4, 5 | $1\ \frac{1}{4}\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{32}$ | $\frac{1}{3}\frac{1}{5}\frac{1}{3}\frac{2}{3} = \frac{2}{135}$ | $\frac{1}{2160}$ |
| 1, 2, 4, 4, 4, 5 | $1\ \frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{16}$ | $\frac{1}{3}\frac{2}{3}\frac{1}{3}\frac{2}{3} = \frac{4}{81}$ | $\frac{1}{324}$ |
| Probabilidad Total | | | $\sum = \frac{77}{10800} \approx 0{,}007$ |

Tabla 1. Probabilidad para todos los caminos permitidos para una secuencia de 4 emisiones en el ejemplo de la Figura 3. Cuando se habla de caminos permitidos se hace referencia a aquellos caminos que no involucren una probabilidad nula.

## 4. La secuencia más probable

En la mayoría de los casos es suficiente con encontrar sólo la mejor secuencia y su probabilidad asociada. Con este fin, existen algoritmos que permiten ahorrar muchos cálculos y, entre ellos, uno de los más utilizados

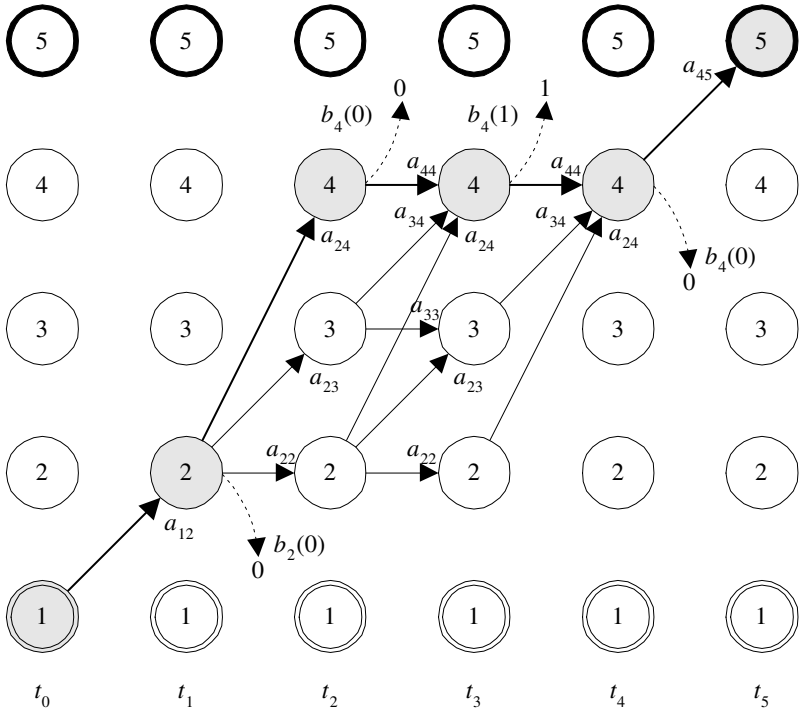


Figura 4. Diagrama de transiciones de estado para el modelo la Figura 3 y una secuencia de 4 observaciones. En este diagrama se indican todos los caminos posibles y se destaca el camino más probable encontrado mediante el algoritmo de Viterbi.

es el algoritmo de Viterbi. En este algoritmo la idea central es recorrer el diagrama de transiciones de estados a través del tiempo, almacenando para cada estado solamente la máxima probabilidad acumulada y el estado anterior desde el que se llega con esta probabilidad. La máxima probabilidad acumulada se obtiene multiplicando la probabilidad de observación del estado por la máxima probabilidad acumulada entre todos los caminos que llegan hasta él. Se entenderá mejor como funciona este algoritmo de definición recursiva mediante un ejemplo.

Para este ejemplo se seguirá el diagrama de la Figura 4, sin olvidar que la secuencia de salida deseada es 0, 0, 1, 0. Se comienza en el estado 1, asignando una probabilidad acumulada $p_1 = 1$ y al pasar al estado 2 la probabilidad acumulada es:

$$p_{12} = b_2(0)\,[p_1 a_{12}] = \frac{1}{3}[1 \times 1] = \frac{1}{3}$$

Desde el estado 2 se puede pasar al 2, al 3 o al 4 obteniendo:

$$p_{122} = b_2(0)\,[p_{12}a_{22}] = \frac{1}{3}\left[\frac{1}{3}\frac{1}{4}\right] = \frac{1}{36}$$

$$p_{123} = b_3(0)\,[p_{12}a_{23}] = \frac{1}{5}\left[\frac{1}{3}\frac{1}{4}\right] = \frac{1}{60}$$

$$p_{124} = b_4(0)\,[p_{12}a_{24}] = \frac{2}{3}\left[\frac{1}{3}\frac{1}{2}\right] = \frac{1}{9}$$

Desde el estado 2 en el tiempo $t_2$ se puede pasar a los estados 2, 3, y 4:

$$p_{1222} = b_2(1)\,[p_{122}a_{22}] = \frac{2}{3}\left[\frac{1}{36}\frac{1}{4}\right] = \frac{1}{216}$$

$$p_{1223} = b_3(1)\,[p_{122}a_{23}] = \frac{4}{5}\left[\frac{1}{36}\frac{1}{4}\right] = \frac{1}{180}$$

$$p_{1224} = b_4(1)\,[p_{122}a_{24}] = \frac{1}{3}\left[\frac{1}{36}\frac{1}{2}\right] = \frac{1}{216}$$

Desde el estado 3 en tiempo $t_2$ se puede pasar a los estados 3 y 4:

$$p_{1233} = b_3(1)\,[p_{123}a_{33}] = \frac{4}{5}\left[\frac{1}{60}\frac{1}{2}\right] = \frac{1}{600}$$

$$p_{1234} = b_4(1)\,[p_{123}a_{34}] = \frac{1}{3}\left[\frac{1}{60}\frac{1}{2}\right] = \frac{1}{360},$$

y desde el estado 4 en el tiempo $t_2$ sólo se puede pasar al estado 4:

$$p_{1244} = b_4(1)\,[p_{124}a_{44}] = \frac{1}{3}\left[\frac{1}{9}\frac{1}{2}\right] = \frac{1}{54}$$

Habiendo llegado al tiempo $t_3$, a partir de cualquiera de los estados solamente es posible pasar al estado 4:

$$p_{12224} = b_4(0)\,[p_{1222}a_{24}] = \frac{1}{3}\left[\frac{1}{216}\frac{1}{2}\right] = \frac{1}{1296}$$

$$\begin{aligned} p_{12?34} &= b_4(0)\,\text{máx}\,\{[p_{1223}a_{34}]\,[p_{1233}a_{34}]\} \\ &= b_4(0)\,\text{máx}\,\{p_{1223}, p_{1233}\}\,a_{34} \\ &= \frac{1}{5}\,\text{máx}\left\{\frac{1}{216}, \frac{1}{600}\right\}\frac{1}{2} \\ &= \frac{1}{5}\frac{1}{216}\frac{1}{2} = \frac{1}{2160} \\ &= p_{12234} \end{aligned}$$

$$\begin{aligned} p_{12?44} &= b_4(0)\,\text{máx}\,\{[p_{1224}a_{44}]\,[p_{1234}a_{44}]\,[p_{1244}a_{44}]\} \\ &= b_4(0)\,\text{máx}\,\{p_{1224}, p_{1234}, p_{1244}\}\,a_{44} \\ &= \frac{2}{3}\,\text{máx}\left\{\frac{1}{216}, \frac{1}{360}, \frac{1}{54}\right\}\frac{1}{2} \\ &= \frac{1}{5}\frac{1}{54}\frac{1}{2} = \frac{1}{162} \\ &= p_{12444} \end{aligned}$$

Finalmente, ya en el tiempo $t_4$ la única opción es pasar al estado 5 que, al igual que el estado 1, es no emisor (ver Figura 3) y no es necesario considerar la probabilidad de observación:

$$\begin{aligned} p_{12??45} &= \text{máx}\,\{[p_{12224}a_{45}]\,[p_{12234}a_{45}]\,[p_{12444}a_{45}]\} \\ &= \text{máx}\,\{p_{12224}, p_{12234}, p_{12444}\}\,a_{45} \\ &= \text{máx}\left\{\frac{1}{1296}, \frac{1}{2160}, \frac{1}{162}\right\}\frac{1}{2} \\ &= \frac{1}{162}\frac{1}{2} = \frac{1}{324} \\ &= p_{124445} \end{aligned}$$

Así, se arriba a la misma conclusión que en el análisis exhaustivo de la Tabla 1: de todos los caminos posibles la mejor secuencia de estados es la 1, 2, 4, 4, 4, 5 y posee una probabilidad de 1/324.

Como se puede observar, se ha ahorrado un gran número de cálculos con este método. En la búsqueda exhaustiva de la Tabla 1 se realizaron 48 multiplicaciones mientras que en el ejemplo de Viterbi sólo fueron 27. Además hay que notar que esta diferencia se incrementa notablemente cuando aumenta el número de estados emisores o la cantidad de observaciones. Esto es debido a que, gracias a que sólo se sigue adelante por los caminos que tienen máxima probabilidad, muchos caminos no se analizan. Se puede ver en este ejemplo que a partir del estado 4 y el tiempo $t_3$, los caminos 1, 2, 2, 4, ?, ? y 1, 2, 3, 4, ?, ? ya no se analizan. Si se conoce una buena forma de llegar a ese estado, solamente se utilizará esta forma. Esto no implica que se deje de lado la evaluación de alguno de los caminos que deriva del estado en cuestión y así el método ahorra muchos cálculos sin perder generalidad.

## 5. Estimación de los parámetros del modelo

Hay que notar que ha quedado de lado una cuestión importante: ¿Cómo se estiman las probabilidades de transición y observación que mejor modelan un conjunto dado de secuencias observadas? Una forma muy intuitiva de entender el entrenamiento es pensar que, si el algoritmo de Viterbi provee la secuencia de estados más probable para una secuencia de símbolos de salida observada, entonces es posible estimar las probabilidades de transición y observación a partir de los símbolos que han quedado asignados a cada estado. Si se posee un conjunto de secuencias observadas para el entrenamiento, se pueden encontrar todas las secuencias de estados más probables y contabilizar las veces que se ha pasado al estado $j$ a partir del estado $i$. A partir de estas cuentas es posible obtener una buena estimación de la probabilidad de pasar al estado $a_{ij}$.

De forma similar, a partir de las secuencias más probables encontradas con el algoritmo de Viterbi, se puede contar la cantidad de veces que el $k$-ésimo símbolo observable ha sido asignado al $j$-ésimo estado del modelo. Esta cuenta puede ser utilizada para obtener una buena estimación de la probabilidad de que el $j$-ésimo estado del modelo emita el $k$-ésimo símbolo observable, es decir, $b_j(k)$.

Mediante una aplicación repetitiva de la búsqueda de la mejor secuencia y posterior reestimación de las probabilidades es posible entrenar el modelo, dado un conjunto de secuencias observadas. Inicialmente se pueden consi-

derar iguales probabilidades para todas las transiciones posibles hacia un estado. De forma similar se pueden considerar inicialmente iguales probabilidades de observación para todos los estados, obtenidas a partir de la cantidad de veces que aparece cada símbolo en el conjunto de secuencias de entrenamiento.

Este método de búsqueda y reestimación se conoce como algoritmo de entrenamiento de Viterbi y es muy rápido en la práctica. Sin embargo, cuando se aplica el algoritmo de Viterbi se trabaja sobre una aproximación de la probabilidad del modelo para cada símbolo de cada secuencia (se ha reemplazado la sumatoria por el máximo). Es así como se obtiene la pertenencia de un símbolo observado a un estado como una función que sólo puede valer 1 o 0 (el símbolo corresponde al estado en cuestión o no corresponde). Si se utiliza una mejor estimación de esta probabilidad, es posible obtener un función de pertenencia con salida no binaria y utilizarla para pesar las evidencias de las secuencias de entrenamiento en la reestimación de las probabilidades del modelo. Éste es el algoritmo de reestimación de Baum-Welch.

## 6. Modelado acústico de la voz

Para seguir aproximando las ideas de MOM al RAH se estudiará cómo utilizarlos para modelar una emisión acústica. Un modelo como el de la Figura 3 podría utilizarse para modelar un fonema y en RAH se denomina *modelo acústico* (MA). Sin embargo, hay que tener en cuenta que los MOM tal como se presentaron hasta el momento, sólo pueden modelar secuencias discretas de símbolos. Esto implica dos niveles de discretización. Por un lado se requiere que los sucesos en el tiempo ocurran a intervalos discretos. Por otro lado se requiere que las manifestaciones de dichos sucesos estén dentro de un conjunto finito de símbolos.

La restricción relativa a la discretización del tiempo puede verse fácilmente superada si se considera el análisis por tramos de la voz ya digitalizada. De esta forma, las observaciones del fenómeno se dan a intervalos regulares de tiempo. En cuanto a la necesidad de que las observaciones pertenezcan a un conjunto finito de símbolos, existen dos posibles alternativas: 1) representar todos los tramos de voz similares mediante un único símbolo y 2) modificar el modelo para que permita modelar valores continuos en las observaciones.

Si se opta por la primera alternativa, luego de dividir la emisión de voz en tramos se busca un símbolo que represente a cada uno. Este proceso

suele incluirse en el denominado *pre-procesamiento* de la señal de voz. Básicamente, una primera etapa del pre-procesamiento se encarga de obtener una representación adecuada del tramo mediante, por ejemplo, un análisis en frecuencia[4]. Luego, una segunda etapa clasifica el tramo de análisis y le asocia uno de los símbolos con que trabaja el MOM. Esta clasificación también puede entenderse como una cuantización, donde un grupo de valores reales se convierte en un número entero dentro de un rango acotado. En la Figura 5 se observan las etapas principales y las señales involucradas. En primer lugar está la señal de voz y luego se esquematiza el análisis por tramos en el tiempo. A continuación cada segmento se analiza en el dominio de la frecuencia y finalmente se realiza una clasificación o cuantización vectorial que da por resultado una secuencia de elementos discretos.

Si se posee un MOM para cada una de las unidades acústicas a modelar (en general fonemas, sílabas o palabras), entonces se podrá aplicar el algoritmo de Viterbi y obtener el mejor camino de cada MOM. Finalmente, el MOM cuyo mejor camino presente la mayor probabilidad será el que determine de qué unidad acústica se trataba.

El esquema que hasta aquí se presenta es el que se conoce como MOM *discreto*, debido a que lo que se modela realmente es una secuencia de símbolos discretos a través de probabilidades de observación discretas. Volviendo a la segunda alternativa para solucionar estas restricciones, se elimina la etapa de cuantización vectorial y se definen los modelos ocultos de Markov *continuos* (MOMC), que utilizan directamente los vectores procedentes del análisis en frecuencia de los tramos de voz. Para esto es necesario replantear las probabilidades de observación de cada estado como, por ejemplo, vectores que contienen las medias y desviaciones para cada elemento del segmento de voz que modelan[5]. De esta manera cada estado de cada modelo tendría sus propias distribuciones de probabilidad que modelan las características acústicas de la voz. Finalmente, existe una alternativa intermedia denominada modelos ocultos de Markov *semicontinuos* (MOMSC), en donde todos los modelos comparten un conjunto fijo de distribuciones de probabilidad.

[4]Muchas de las características que permiten una clasificación de los sonidos del habla se hacen evidentes en el dominio de la frecuencia.

[5]Este es un ejemplo muy simplificado y tienen por finalidad transmitir solamente el concepto general de modelados mediante MOMC.

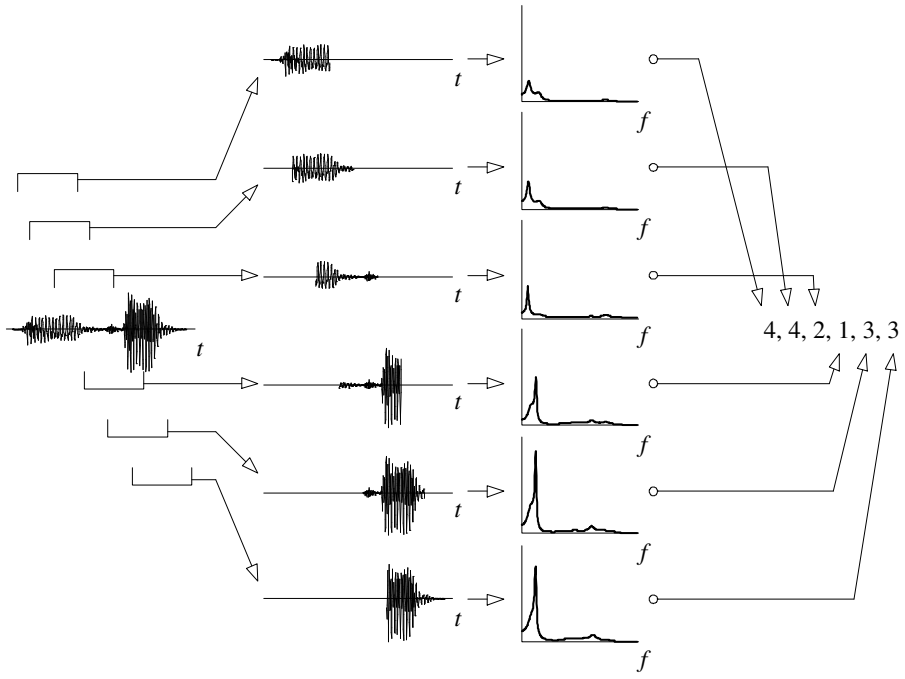


Figura 5. Procesamiento necesario para utilizar modelos ocultos de Markov discretos en reconocimiento automático del habla. Se pueden observar las etapas principales y las señales involucradas. En primer lugar está la señal de voz y luego se esquematiza el análisis por tramos. A continuación cada segmento se analiza en el dominio de la frecuencia y finalmente se realiza una clasificación que da por resultado una secuencia de elementos discretos. Los modelos ocultos de Markov continuos no requieren esta última etapa y trabajan directamente con los vectores en el dominio transformado.

## 7. El modelo de lenguaje y el modelo compuesto

Cuando se habla del modelo de lenguaje (ML), se sitúa el estudio en niveles superiores al de las características acústicas, por encima de los fonemas y los suprasegmentos. Ahora interesan las palabras y la forma en que se combinan para formar frases. Siguiendo con la idea de los autómatas probabilísticos (finitos), es posible imaginar un autómata en el que cada estado represente (o emita) una palabra. En la Figura 6 se puede observar una estructura que respeta la idea general de un autómata probabilístico como el de la Figura 2 (página 6), utilizado para modelar secuencias temporales de palabras. Estas estructuras son conocidas como *gramáticas* en la teoría de

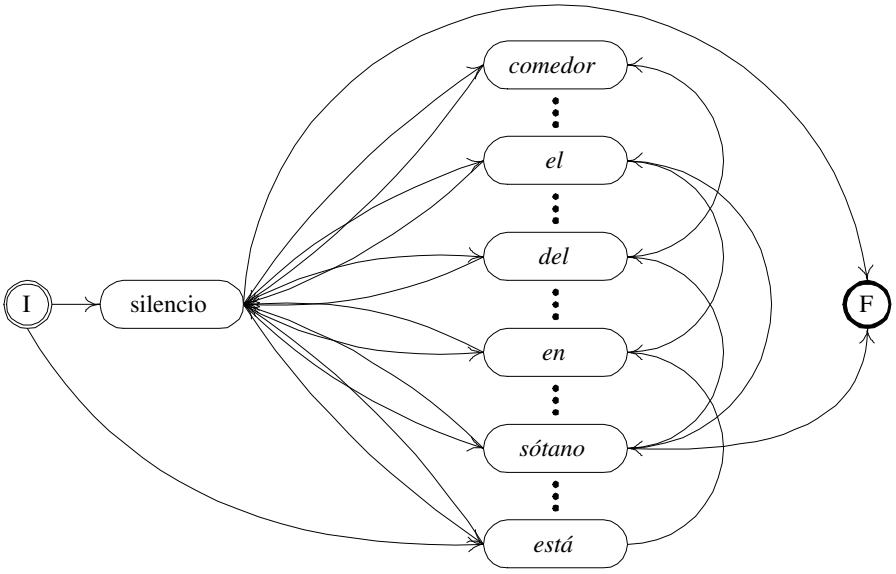


Figura 6. Modelo de lenguaje. Los estados de inicio y finalización se indican con las letras I y F, respectivamente.

lenguajes formales y conservan ese nombre en la jerga del RAH.

Sin embargo, se puede observar que la secuencia de estados de una de estas gramáticas es también una cadena de Markov y así se pueden extender los formalismos de los MOM para incluir estas representaciones en un nivel superior al acústico. A partir de una descripción fonética de cada palabra, conocida como *diccionario fonético*, se podrían formar las palabras de este ML concatenando los MA de los diferentes fonemas. Finalmente se construiría un *modelo compuesto* (MC) capaz de modelar cualquier frase, desde los aspectos fonéticos más elementales hasta las complejidades del lenguaje hablado. En la Figura 7 se pueden observar los tres niveles de la composición: el ML, el diccionario fonético y el MA.

Mediante este MC es posible formar modelos para diferentes frases y evaluar, con una extensión del algoritmo de Viterbi, las probabilidades de cada frase para una emisión de voz dada. El proceso de reconocimiento culmina eligiendo el modelo de la frase que mayor probabilidad posea y dando como resultado el texto con que se formó la frase. Cabe aclarar que, nuevamente, la búsqueda sobre todas las frases posibles no se realiza de forma exhaustiva. Para esto existe una gran variedad de algoritmos que organizan y recorren de diferentes formas la expansión del MC.

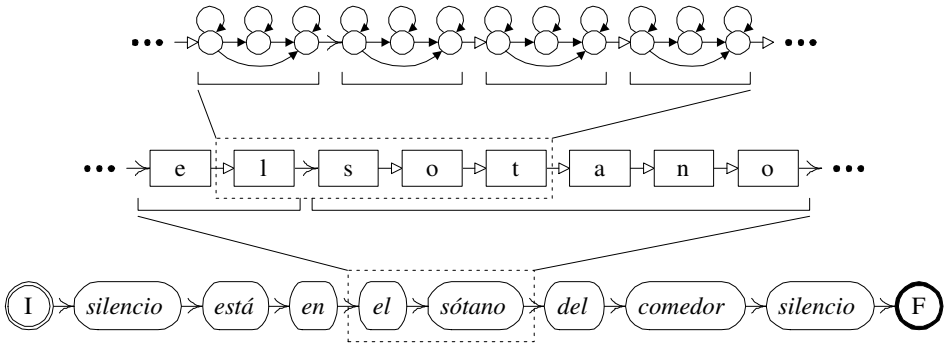


Figura 7. Modelo compuesto para la frase: *Está en el sótano del comedor*. Se pueden observar los tres niveles de la composición: los estados del modelo acústico, el diccionario fonético y el modelo de lenguaje. En los modelos acústicos se han eliminado los estados no emisores para simplificar el esquema.

Resta por comentar brevemente la extensión de los algoritmos de entrenamiento para el MC. Existen dos conjuntos de parámetros a estimar durante el entrenamiento: las probabilidades de transición y observación de los MA y las probabilidades de transición del ML. Estas estimaciones se realizan separadamente, es decir, se estiman las primeras dejando fijo el ML y viceversa. Para la estimación de las probabilidades de los MA, a partir de una de las frases de entrenamiento y dada su transcripción en texto es posible formar un MC para esta frase y luego aplicar el algoritmo de entrenamiento sobre este gran modelo, tal como se aplicó en el caso de un pequeño MOM. Los mismos modelos de fonemas o palabras pueden concatenarse para formar otro MC de frase y nuevamente realizar un ajuste mediante el algoritmo de entrenamiento. Las probabilidades que corresponden al ML, que habían quedado fijas durante este proceso, son estimadas directamente del texto de las frases de entrenamiento, contando la cantidad de veces que aparece una determinada secuencia y asignado una probabilidad a las transiciones que es proporcional a esta cuenta.

# 8. Bibliografía básica

[Deller et al., 1993] Deller, J. R., Proakis, J. G., y Hansen, J. H. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing, NewYork.

[Ferguson, 1980] Ferguson, J. *Hidden Markov Models for Speech*. IDA, Princeton, NJ.

[Huang et al., 1990] Huang, X. D., Ariki, Y., y Jack, M. A. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press.

[Jelinek, 1999] Jelinek, F. *Statistical Methods for Speech Recognition*. MIT Press, Cambrige, Masachussets.

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