# Fundamentos del reconocimiento automático del habla

Sugerencias y correcciones a:
d.milone@ieee.org
29 de febrero de 2004

En este documento se hará una descripción detallada de las principales técnicas utilizadas en el reconocimiento automático del habla (RAH). El trabajo se divide en dos grandes partes: el análisis de la señal de voz y los modelos ocultos de Markov. En primer lugar se tratará, como marco general, el análisis por tramos de la señal de voz. A partir de esta particular forma de seguir la dinámica de la voz, se describen los diferentes métodos de análisis. En la segunda parte se describe la estructura y entrenamiento de un sistema de reconocimiento automático del habla basado en modelos ocultos de Markov. Inicialmente se trata en forma genérica la versión continua de estos modelos y luego se realiza una ampliación para incluir a los modelos semicontinuos. Para completar esta descripción se incluyen los modelos de palabra y los modelos de lenguaje, construyendo así un modelo compuesto. Las ecuaciones para el entrenamiento y la decodificación se extienden al modelo compuesto utilizado en el reconocimiento automático del habla continua[1].

# 1. Análisis de la señal de voz

La señal de voz posee una gran variabilidad en el tiempo y es necesario descomponerla en intervalos que permitan su estudio bajo la hipótesis de estacionariedad. Estos intervalos de tiempo estarán en relación directa con la máxima velocidad con que el tracto vocal pueda modificar significativamente su morfología. En las aplicaciones prácticas para el RAH se utilizan intervalos de 10 a 30 ms. A continuación se desarrollan estas ideas y a partir de allí se definen técnicas útiles para el análisis de la señal de voz en el contexto del RAH.

## 1.1. Análisis por tramos

Sea $v(\tau)$ la señal continua de voz para la variable real de tiempo $\tau$. Después de un proceso de muestreo uniforme con período $T_v$, la señal de voz en la variable natural de tiempo discreto $0 < m \leq N_v$ se representa como $v(mT_v)$ o más simplemente $v(m)$.

Sea la señal $\omega(m; N_\omega)$ una ventana de análisis definida para $0 < m \leq N_\omega$, se dice que esta ventana posee un *ancho* $T_\omega = N_\omega T_v$. De la aplicación de la ventana de análisis temporal se obtienen los *tramos* de voz:

[1] Para un tratamiento introductorio y conceptual se ha escrito otro documento "Modelos ocultos de Markov para el reconocimiento automático del habla: una breve introducción".

$$v(t;n) = \omega(n;N_\omega)v(tN_d + n); \quad 0 < n \leq N_\omega \tag{1}$$

que representaremos en notación vectorial como $\mathbf{v}_t$. Se denomina *paso* del análisis por tramos al tiempo $T_d = N_d T_v$. Dadas las definiciones anteriores la variable de tiempo por tramos $t \in \mathbb{N}$ queda acotada según $0 < t \leq T = (N_v - N_\omega)/N_d + 1 < \infty$.

Si $\mathcal{T}(k)$ es un operador para la transformación de dominio, se realiza el proceso de parametrización de la señal de voz según:

$$x(t;k) = \mathcal{T}(k)\left\{v(t;n)\right\}, \quad 0 < k \leq N_x$$

para la que se utilizará la notación vectorial simplificada como $\mathbf{x}_t \in \mathbb{X} = \mathbb{R}^{N_x}$. Se conoce a $\mathbb{X}$ como el *espacio de las características* con dimensión $N_x$. En esta sección se utilizará $0 < k \leq N_x$ como variable independiente discreta en el dominio transformado.

### 1.1.1. Ventanas de análisis

Las ventanas de análisis más utilizadas se definen para $0 < m \leq N_\omega$ según:

I) Ventana rectangular:

$$\omega_R(m;N_\omega) = 1$$

II) Ventana de Hanning:

$$\omega_h(m;N_\omega) = \frac{1}{2} - \frac{1}{2}\cos(2\pi m/N_\omega)$$

III) Ventana de Hamming:

$$\omega_H(m;N_\omega) = \frac{27}{50} - \frac{23}{50}\cos(2\pi m/N_\omega)$$

IV) Ventana de Bartlett:

$$\omega_B(m; N_\omega) = \begin{cases} 2m/N_\omega & \text{si } 0 < m \leq N_\omega/2 \\ 2 - 2m/N_\omega & \text{si } N_\omega/2 < m \leq N_\omega \end{cases}$$

V) Ventana de Blackman:

$$\omega_K(m; N_\omega) = \frac{21}{50} - \frac{1}{2}\cos(2\pi m/N_\omega) + \frac{2}{25}\cos(4\pi m/N_\omega)$$

Estas ventanas pueden ser caracterizadas por el tamaño de los lóbulos de la magnitud de su espectro de frecuencias. La ventana rectangular posee el lóbulo central con menor ancho de banda pero la magnitud de los lóbulos laterales decae muy lentamente. La ventana de Blackman posee la mínima amplitud en sus lóbulos laterales pero su lóbulo principal tiene un ancho de banda tres veces mayor al de la rectangular [Kuc, 1988]. Dado este compromiso entre resolución frecuencial y distorsión armónica en el proceso de ventaneo, para señales de voz suele utilizarse la ventana de Hamming que además, ofrece una posición media en cuanto al costo computacional de su aplicación [Deller et al., 1993].

### 1.1.2. Transformaciones

El operador $\mathcal{T}(k)$ permite obtener un vector de características $\mathbf{x}_t$ para el análisis por tramos de la señal de voz. A continuación se tratarán los operadores más comúnmente utilizados en el RAH:

I) Coeficientes espectrales (CE):

$$\mathbf{x}_t = [u(t;k)] = \mathcal{T}_F(k)\left\{v(t;n)\right\},$$

II) Coeficientes de predicción lineal (CPL):

$$\mathbf{x}_t = [a(t;k)] = \mathcal{T}_L(k)\left\{v(t;n)\right\},$$

III) Coeficientes cepstrales (CC):

$$\mathbf{x}_t = [c(t;k)] = \mathcal{T}_C(k)\left\{v(t;n)\right\},$$

En las diferentes alternativas para los vectores de características se definirán $N_{uI}$, $N_a$ y $N_{cI}$ que, en el caso general, corresponderán a $N_x$.

## 1.2. Coeficientes espectrales

Se define la transformada discreta de Fourier (TDF) de $v(m)$ como:

$$u(k) = \sum_{m=1}^{N_v} v(m) e^{-\mathrm{j}(2\pi/N_v)k(m-1)} \tag{2}$$

Si se aplica la TDF a los tramos de voz $v(t;n)$ de la ecuación (1), es posible obtener la denominada transformada de Fourier de tiempo corto o por tramos:

$$\begin{aligned} u(t;k) &= \sum_{n=1}^{N_v} v(t;n) e^{-\mathrm{j}(2\pi/N_v)(k-1)(n-1)} \\ &= \sum_{n=1}^{N_v} \omega(n;N_v) v(tN_d+n) e^{-\mathrm{j}(2\pi/N_v)(k-1)(n-1)} \end{aligned}$$

Generalmente, dado que $v(t;n) \in \mathbb{R}$, se utiliza el espectro de magnitud $|u(t;k)|$ en $0 < k \leq N_v/2$ y la notación vectorial $\mathbf{u}_t \in \mathbb{R}^{N_u}$ con $N_u = N_v/2$.

### 1.2.1. Integración por bandas

Para el RAH suele utilizarse el logaritmo de la energía de un número reducido de bandas del espectro, en lugar del espectro completo. Es necesario definir las frecuencias de corte para cada banda y para cada ley de mapeo frecuencial o "escala" de integración se podrá obtener un conjunto diferente de coeficientes. Un ejemplo sencillo es la escala de integración lineal, donde la relación entre ambas frecuencias tiene la forma:

$$F_{lin} \propto f_{Hz}$$

Si se consideran $N_{uI}$ bandas de integración en la primera mitad del espectro, es posible calcular los extremos de cada intervalo mediante:

$$B(k) = \frac{kN_u}{2N_{uI}}; \quad 0 \leq k \leq N_{uI}$$

En el caso más simple se realiza la integración mediante ventanas frecuenciales rectangulares:

$$u_I(t;k) = 2 \sum_{\varkappa=B(k-1)}^{\varkappa=B(k)} \log|u(t;\varkappa)|; \quad 0 < k \leq N_{uI}$$

Cuando se utilizan ventanas de Bartlett o de Hamming el esquema de integración se modifica para no perder la energía en los extremos de cada ventana:

$$u_I(t;k) = 2 \sum_{\varkappa=B(k-1)}^{\varkappa=B(k+1)} \omega_B\left(\varkappa - B(k-1); B(k+1) - B(k-1)\right) \log|u(t;\varkappa)| \tag{3}$$

con $0 < k < N_{uI}$.

Diversos estudios acerca de la percepción de tonos puros en el ser humano [Stevens, 1998] han permitido aproximar la relación entre la frecuencia percibida y la real mediante:

$$F_{mel}\left(f_{Hz}\right) = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right),$$

relación que es ampliamente utilizada como escala de integración en el procesamiento de señales de voz.

## 1.3. Coeficientes de predicción lineal

Es posible modelar el tracto vocal mediante un sistema auto-regresivo de la forma:

$$\hat{v}(t;n) = -\sum_{j=1}^{N_a} a(t;j)v(t;n-j) + Gg(t;n) \tag{4}$$

donde $v(t;n)$ es la señal a modelar, $\hat{v}(t;n)$ es la señal estimada por el modelo, $g(t;n)$ es la entrada al tracto vocal y $N_a$ es el orden del sistema.

Para este análisis se considera inicialmente una entrada nula y la ecuación anterior puede escribirse usando notación vectorial simplificada como:

$$\hat{v}(t;n) = -(\mathbf{v}_t^n)^{\mathrm{T}}\mathbf{a}_t$$

donde $\mathbf{a}_t$ contiene los $N_a$ coeficientes $a(t;j)$ y $\mathbf{v}_t^n$ contiene las últimas $N_a$ salidas $v(t;n-j)$. El error entre $v(t;n)$ y $\hat{v}(t;n)$ se puede medir mediante la distancia euclídea como:

$$e(t;n)^2 = (v(t;n) - \hat{v}(t;n))^2.$$

Para encontrar el vector $\mathbf{a}_t$ se minimiza la medida del error cuadrático total entre $\hat{v}(t;n)$ y $v(t;n)$:

$$\xi^2 = \sum_n e(t;n)^2 = \sum_n \left(v(t;n) + (\mathbf{v}_t^n)^{\mathrm{T}}\mathbf{a}_t\right)^2$$

haciendo:

$$\nabla\xi^2 = 0$$

se obtiene:

$$\left(\sum_n \mathbf{v}_t^n(\mathbf{v}_t^n)^{\mathrm{T}}\right)\mathbf{a}_t = -\sum_n \mathbf{v}_t^n v(t;n)$$

conocido como sistema de Wiener-Hopf y comúnmente representado como:

$$\mathbf{R}_t \mathbf{a}_t = -\mathbf{r}_t \tag{5}$$

donde $\mathbf{r}_t$ es el vector de autocorrelación y $\mathbf{R}_t$ la matriz de autocorrelación para $v(t;n)$. Se puede verificar que $R_{ij} = r_{i-j}$ y así $\mathbf{R}_t$ es una matriz Toeplitz. El método de Levinson-Durbin [Kay y Marple, 1981] aprovecha esta propiedad para simplificar la resolución del sistema.

Resta por definir el orden $N_a$ del sistema. Existen varios métodos para encontrar el orden del sistema de forma que se obtenga un buen compromiso entre el error total y la complejidad de su estructura. Estos métodos se basan en medidas del error en la predicción, por ejemplo, a partir de las ecuaciones (4) y (5) es posible obtener ([Makhoul, 1975]):

$$E(N_a) = r_0 + \mathbf{r}_t^{\mathrm{T}} \mathbf{a}_t$$

y encontrando el modelo más simple cuyo $E(N_a)$ sea mínimo se puede determinar el orden apropiado para la estimación. Otros métodos más elaborados [Akaike, 1974] utilizan criterios basados en la teoría de la información. Suponiendo una distribución gaussiana en la señal se puede medir el error según:

$$I(N_a) = \log E(N_a) + \frac{2N_a}{N_e}$$

donde $N_e$ es el número efectivo de muestras en la señal, que para el caso de una ventana de Hamming $N_e = 0{,}4N_\omega$. En general, para el modelado de señales de voz en RAH se encuentra un buen compromiso para un orden $N_a$ entre 10 y 16 [Young et al., 2000].

## 1.4. Coeficientes cepstrales

En base a la TDF, se define el cepstrum real de $v(m)$ como:

$$c(m) = \mathcal{T}_F^{-1} \left\{ \log \left| \mathcal{T}_F \left\{ v(m) \right\} \right| \right\},$$

Esta definición se puede extender para un análisis por tramos. Reemplazando según la TDF (2) y su inversa (TDFI), se obtiene:

$$\begin{aligned} c(t;k) &= \mathcal{T}_F^{-1}\left\{\log\left|\sum_{n=1}^{N_v} v(t;n)e^{-j(2\pi/N_v)(\varkappa-1)(n-1)}\right|\right\} \\ &= \frac{1}{N_u}\sum_{\varkappa=1}^{N_u}\log|u(t;\varkappa)|\, e^{j(2\pi/N_u)(\varkappa-1)(k-1)} \end{aligned} \tag{6}$$

Finalmente, si se considera que el argumento de la TDFI es una secuencia real y par, puede simplificarse su cómputo mediante una transformada coseno (TC):

$$\begin{aligned} c(t;k) &= \frac{1}{N_u}\sum_{\varkappa=1}^{N_u}\log|u(t;\varkappa)|\cos\left((2\pi/N_u)(\varkappa-1)(k-1)\right) \\ &= \frac{2}{N_u}\sum_{\varkappa=2}^{N_u/2-1}\log|u(t;\varkappa)|\cos\left((2\pi/N_u)(\varkappa-1)(k-1)\right) \end{aligned} \tag{7}$$

### 1.4.1. La señal de voz y el cepstrum

Siguiendo la idea del modelo para el tracto vocal presentada en la ecuación (4), es posible considerar que la señal de voz para fonemas sonoros es generada mediante la convolución:

$$\hat{v}(t;n) = g(t;n) * h(t;n)$$

donde la entrada al sistema es el tren de pulsos glóticos $g(t;n)$ y $h(t;n)$ es la respuesta al impulso del tracto vocal. Cuando se pasa al dominio frecuencial mediante $\mathcal{T}_F$ y se aplica el logaritmo, resulta:

$$\hat{v}(t;\varkappa) = \log|g(t;\varkappa)| + \log|h(t;\varkappa)|$$

Cuando nuevamente se transforma esta señal mediante la TDFI se obtiene:

$$\hat{v}(t;k) = \mathcal{T}_F^{-1}\left\{\log|g(t;\varkappa)|\right\} + \mathcal{T}_F^{-1}\left\{\log|h(t;\varkappa)|\right\}$$

Generalmente, la señal del pulso glótico varía muy lentamente en relación a la otra componente, digamos, con período $1/F_0$. Cuando se realiza la primera transformación, claramente se puede observar que $g(t;\varkappa)$ es modulada por $h(\varkappa)$ a razón de $F_0$. Es así como la segunda transformación, luego de haber aplicado el logaritmo a la magnitud, deja en las primeras muestras la información relacionada con $h(t;k)$ y a partir de $1/F_0$ lo relativo al pulso glótico $g(t;k)$. Normalmente, en RAH se utiliza la primera parte del cepstrum y se descarta lo relativo al pulso gótico.

En base a la información de la segunda parte del cepstrum se han descrito muchos métodos para estimar la frecuencia fundamental ($F_0$) en señales de voz [Hess, 1991]. Además de aquellos basados en CC, existen métodos basados en la correlación cruzada, en CE y en CPL [Deller et al., 1993]. Siguiendo el razonamiento anterior, en relación a la forma en que el cepstrum real separa la información relativa al pulso glótico, se puede observar que la simple detección del pico correspondiente al pulso glótico en el cepstrum real constituye un método para determinar $F_0$ en los fonemas sonoros. Los estudios en este sentido fueron iniciados por Michael Noll, quien reunió un conjunto de reglas sencillas para eliminar los principales artefactos generados al aplicar el método en voz continua [Noll, 1967]. Este conjunto reducido de reglas aplicadas a los CC sigue siendo hasta la actualidad el mejor método conocido para la determinación de la entonación en habla limpia [Shimamura y Kobayashi, 2001].

### 1.4.2. Coeficientes cepstrales en escala de mel

Para combinar las propiedades del cepstrum y los resultados acerca de la percepción de tonos puros en el ser humano, se propuso integrar la representación espectral de la señal según la escala de mel antes de aplicar la TC [Davis y Mermelstein, 1980]. Siguiendo estas ideas se pueden definir los coeficientes cepstrales en escala de mel (CCEM) a partir de las ecuaciones (3) y (6):

$$
\begin{aligned}
c_{mel}(t;k) &= \frac{2}{N_{uI}} \sum_{i=2}^{N_{uI}} u_I(t;i) \cos\left((2\pi/N_{uI})(i-1)(k-1)\right) \\
&= \frac{4}{N_{uI}} \sum_{i=2}^{N_{uI}} \sum_{\varkappa=B(i-1)}^{\varkappa=B(i+1)} \omega_B\left(\varkappa - B(i-1); B(i+1) - B(i-1)\right) \\
&\qquad \times \log \left| \sum_{n=1}^{N_v} v(t;n) e^{-j(2\pi/N_v)(\varkappa-1)(n-1)} \right| \\
&\qquad \times \cos\left((2\pi/N_{uI})(i-1)(k-1)\right)
\end{aligned}
$$

Los resultados experimentales han favorecido ampliamente esta combinación. Como detalle de aplicación práctica debe mencionarse que en general para RAH no se utilizan todos los $c_{mel}(t;k)$ sino que se desecha toda la información relacionada con el pulso glótico. De forma similar que para los CPL, suelen utilizarse los primeros $N_{cI} = 13$ CCEM, a partir de una integración según $N_{uI} = 24$ bandas.

### 1.4.3. Relación entre CC y CPL

Para completar lo relativo a CC, se describe a continuación una aproximación que resulta útil para su cálculo [Huang et al., 1990]. Denotando por $\mathcal{Z}\{\cdot\}$ al operador de la transformada Z [Oppenheim y Schafer, 1989], es posible escribir:

$$
\frac{\mathrm{d}\mathcal{Z}\left\{v(t;n)\right\}}{\mathrm{d}z^{-1}} = \mathcal{Z}\left\{v(t;n)\right\} \frac{\mathrm{d}\mathcal{Z}\left\{c(t;k)\right\}}{\mathrm{d}z^{-1}}
$$

ya que $\log\left(\mathcal{Z}\left\{v(t;n)\right\}\right) = \mathcal{Z}\left\{c(t;k)\right\}$. Considerándose que una estimación del espectro de la señal de voz puede obtenerse a partir del modelo autoregresivo de la ecuación (4):

$$
\mathcal{Z}\left\{v(t;n)\right\} \approx \frac{G}{\mathcal{Z}\left\{a(t;j)\right\}}
$$

ahora se obtiene:

$$-\mathcal{Z}\left\{ja(t;j)\right\} \approx \mathcal{Z}\left\{kc(t;k)\right\}\mathcal{Z}\left\{a(t;j)\right\}.$$

Invirtiendo la transformada Z y teniendo en cuenta que el producto del término de la derecha quedará como una convolución en el dominio no transformado:

$$\hat{c}(t;k) = -a(t;k) - \frac{1}{k}\sum_{j=2}^{k}(k-j+1)c(t;k-j+1)a(t;j)$$

con $a(t;i) = 0$ para $i > p$ y $k \geq 2$, ya que de la ecuación (7) se puede ver que $c(t;1) \propto \sum_{\varkappa} \log|u(t;\varkappa)|$.

## 1.5. Coeficientes de energía, delta y aceleración

Cuando se confecciona el vector de características para RAH, es práctica corriente considerar algunas otras variables que llevan información importante del tramo de voz considerado. Una de estas variables consiste en una medida de la energía que se define simplemente como:

$$\epsilon(t) = \log\sum_{n=1}^{N_v} v(t;n)^2 \tag{8}$$

También suele agregarse una estimación de las derivadas temporales de todos los elementos calculados. Para un vector de características $x(t;k)$ dado, se obtienen los *coeficientes delta* mediante la regresión:

$$\Delta x(t;k) = \frac{\sum_{j=1}^{N_J} j\left(x(t+j;k) - x(t-j;k)\right)}{2\sum_{j=1}^{N_J} j^2}$$

donde $N_J$ es utilizado para suavizar la estimación a través de los tramos (generalmente $1 \leq N_J \leq 2$). Los coeficientes de aceleración $\Delta^2 x(t;k)$ se obtienen por aplicación directa de la ecuación anterior a los $\Delta x(t;k)$.

# 2. Modelos ocultos de Markov

En esta sección se tratarán formalmente los modelos ocultos de Markov (MOM). Para comenzar se definirán los MOM continuos (MOMC) y se deducirán las fórmulas para la estimación de sus parámetros. A continuación se harán las extensiones necesarias para cubrir la estructura y el entrenamiento de los MOM semicontinuos. Finalmente se tratarán los modelos de lenguaje y su incorporación en lo que denominamos modelos compuestos para el RAH.

## 2.1. Estructura del modelo

Un MOMC queda definido mediante una estructura algebraica:

$$\Theta = \langle \mathcal{Q}, \mathbb{O}, \mathbf{A}, \mathcal{B} \rangle$$

donde:

- $\mathcal{Q}$ es el conjunto de estados posibles,
- $\mathbb{O}$ es el espacio observable,
- $\mathbf{A}$ es la matriz de probabilidades de transición de estados y
- $\mathcal{B}$ es el conjunto de distribuciones de probabilidades de observación.

El conjunto de estados posibles se define como:

$$\mathcal{Q} = \{q \in [1 \dots |\mathcal{Q}|]\}\,; \quad |\mathcal{Q}| < \infty$$

donde $|\mathcal{Q}| \in \mathbb{N}$ es la cardinalidad del conjunto. Para el espacio observable se tiene:

$$\mathbb{O} = \left\{\mathbf{o} \in \mathbb{R}^{N_o}\right\}; \quad N_o = N_x$$

donde $N_o \in \mathbb{N}$ es su dimensión, que coincide con la dimensión del espacio de las características $\mathbb{X}$, que en el contexto del los MOM también se denominará espacio de las evidencias acústicas.

Sean $q_{t-1}, q_t \in \mathcal{Q}$ dos estados cualquiera de modelo $\Theta$, donde $t \in [1 \dots T] \subset \mathbb{N}$ tal como se definió en la Sección 1.1, entonces se define la matriz de probabilidades de transición de estados como:

$$\mathbf{A} = [a_{ij} = \Pr(q_t = j \,|q_{t-1} = i)] \quad \forall i, j \in \mathcal{Q}$$

donde $a_{ij} \geq 0 \;\; \forall i, j$ y $\sum_{j=1}^{|\mathcal{Q}|} a_{ij} \stackrel{\circ}{=} 1 \;\; \forall i \in \mathcal{Q}$.

Siendo $\mathbf{x}_t \in \mathbb{X}$ una evidencia acústica para el modelo $\Theta$, se define el conjunto de distribuciones de probabilidad de observación como:

$$\mathcal{B} = \{b_j(\mathbf{x}_t) = \Pr(\mathbf{x}_t \,|q_t = j)\} \quad \forall j \in \mathcal{Q}$$

en donde para cada estado $j$ se modela la distribución de probabilidades mediante la mezcla:

$$b_j(\mathbf{x}_t) = \sum_{k=1}^{N_c} c_{jk} b_{jk}(\mathbf{x}_t) \quad \forall j \in \mathcal{Q};\; N_c < \infty \tag{9}$$

siendo en este caso:

I) $b_{jk}(\mathbf{x}_t)$: todas funciones normales de densidad de probabilidad multidimensional con la forma:

$$\mathcal{N}(\mathbf{x}_t, \boldsymbol{\mu}_{jk}, \mathbf{U}_{jk}) = \frac{1}{(2\pi)^{N_x} |\mathbf{U}_{jk}|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(\mathbf{x}_t - \boldsymbol{\mu}_{jk})^{\mathrm{T}} \mathbf{U}_{jk}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_{jk})\right]},$$

II) $c_{jk} \in \mathbb{R}^{+0}$: las constantes de peso relativo para cada distribución normal que satisfacen:

$$\sum_{k=1}^{N_c} c_{jk} \stackrel{\circ}{=} 1 \;\; \forall j \in \mathcal{Q},$$

III) $\boldsymbol{\mu}_{jk} \in \mathbb{R}^{N_x}$: los vectores de medias,

IV) $\mathbf{U}_{jk} \in \mathbb{R}^{N_x \times N_x}$: las matrices de covarianza y

V) se cumple que:

$$\int_{-\infty}^{+\infty} b_j(\mathbf{x}_t)\,\mathrm{d}\mathbf{x}_t \stackrel{\circ}{=} 1 \quad \forall j \in \mathcal{Q}.$$

## 2.2. La secuencia más probable

Dada la secuencia de evidencias acústicas:

$$\mathbf{X}^T = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T; \quad \mathbf{x}_t \in \mathbb{X}$$

y

$$\mathbf{q}^T = q_1, q_2, \ldots, q_T; \quad q_t \in \mathcal{Q}$$

una secuencia cualquiera de exactamente $T$ estados, se calcula la probabilidad de que el modelo $\Theta$ haya generado la secuencia de evidencias acústicas $\mathbf{X}^T$ mediante:

$$\Pr\left(\mathbf{X}^T|\Theta\right) = \sum_{\forall \mathbf{q}^T} \Pr\left(\mathbf{X}^T, \mathbf{q}^T|\Theta\right) \tag{10}$$

Bajo la hipótesis de independencia estadística de las evidencias acústicas en $\mathbf{X}^T$:

$$\begin{aligned}
\Pr\left(\mathbf{X}^T|\Theta\right) &= \sum_{\forall \mathbf{q}^T} \left\{\Pr\left(\mathbf{X}^T|\mathbf{q}^T,\Theta\right)\Pr\left(\mathbf{q}^T|\Theta\right)\right\} \\
&= \sum_{\forall \mathbf{q}^T} \left\{\prod_{t=1}^{T}\Pr\left(\mathbf{x}_t|q_t,\Theta\right)\prod_{t=2}^{T}\Pr\left(q_t|q_{t-1},\Theta\right)\right\} \\
&= \sum_{\forall \mathbf{q}^T} \left\{\prod_{t=1}^{T} b_{q_t}\left(\mathbf{x}_t\right)\prod_{t=2}^{T} a_{q_{t-1}q_t}\right\} \\
&= \sum_{\forall \mathbf{q}^T} \left\{b_{q_1}\left(\mathbf{x}_1\right)\prod_{t=2}^{T} b_{q_t}\left(\mathbf{x}_t\right)a_{q_{t-1}q_t}\right\}
\end{aligned}$$

que puede simplificarse en:

$$\Pr\left(\mathbf{X}^T|\Theta\right) = \sum_{\forall \mathbf{q}^T}\prod_{t=1}^{T} b_{q_t}\left(\mathbf{x}_t\right)a_{q_{t-1}q_t} \tag{11}$$

haciendo $a_{01} = 1$.

Una buena aproximación para $\Pr\left(\mathbf{X}^T|\Theta\right)$ es considerar la función de máximo en lugar de la sumatoria sobre las secuencias $\mathbf{q}^T$:

$$\Pr\left(\mathbf{X}^T|\Theta\right) \approx \max_{\forall \mathbf{q}^T}\left\{\Pr\left(\mathbf{X}^T|\mathbf{q}^T,\Theta\right)\Pr\left(\mathbf{q}^T|\Theta\right)\right\}$$

El algoritmo de Viterbi optimiza la búsqueda de esta máxima probabilidad. Para esto se define la variable de probabilidad acumulada:

$$\lambda_t(j) \triangleq \max_{\forall \mathbf{q}^{t-1}}\left\{\Pr\left(\mathbf{q}^{t-1}, q_t = j, \mathbf{X}^t|\Theta\right)\Pr\left(\mathbf{q}^{t-1}|\Theta\right)\right\};\quad \forall j \in \mathcal{Q} \tag{12}$$

con $\lambda_0(j) = 1\ \forall j \in \mathcal{Q}$, y calculable por inducción mediante la recursión:

$$\begin{aligned}
\lambda_t(j) &= \max_{\forall i \in \mathcal{Q}} \{\lambda_{t-1}(i) \Pr(q_t = j, \mathbf{x}_t | q_{t-1} = i, \Theta)\} \\
&= \max_{\forall i \in \mathcal{Q}} \{\lambda_{t-1}(i) \Pr(q_t = j | q_{t-1} = i, \Theta) \Pr(\mathbf{x}_t | q_t = j, \Theta)\} \\
&= \max_{\forall i \in \mathcal{Q}} \{\lambda_{t-1}(i) \Pr(q_t = j | q_{t-1} = i, \Theta)\} \Pr(\mathbf{x}_t | q_t = j, \Theta) \\
&= \max_{\forall i \in \mathcal{Q}} \{\lambda_{t-1}(i) a_{ij}\} b_j(\mathbf{x}_t) \qquad (13)
\end{aligned}$$

de forma que:

$$\Pr\left(\mathbf{X}^T | \Theta\right) \approx \max_{\forall j \in \mathcal{Q}} \{\lambda_T(j)\}.$$

Para encontrar la secuencia de estados $\tilde{\mathbf{q}}^T$ asociada a la máxima probabilidad se define:

$$\xi_t(j) \triangleq \arg \max_{\forall i \in \mathcal{Q}} \{\lambda_{t-1}(i) a_{ij}\}$$

y a partir de:

$$\tilde{q}_T = \arg \max_{\forall i \in \mathcal{Q}} \{\lambda_T(i)\}$$

por recursión inversa:

$$\tilde{q}_t = \xi_{t+1}(\tilde{q}_{t+1}); \quad t = T-1, T-2, \ldots, 1 \qquad (14)$$

## 2.3. Reestimación de los parámetros

Dada una secuencia de evidencias acústicas $\mathbf{X}^T$, el entrenamiento consiste en maximizar la función de densidad de probabilidad $p(\mathbf{X}^T|\Theta)$, que posee la forma de (11). El método para el entrenamiento se fundamenta en la definición de una función auxiliar que guía el proceso de optimización permitiendo obtener una nueva estimación de los parámetros del modelo a

partir de la estimación anterior. La definición que se utiliza en este caso está basada en la teoría de la información[2] y tiene la siguiente forma:

$$\mathcal{O}(\Theta, \tilde{\Theta}) \triangleq \frac{1}{p(\mathbf{X}^T|\Theta)} \sum_{\forall \mathbf{q}^T} p(\mathbf{X}^T, \mathbf{q}^T|\Theta) \log p(\mathbf{X}^T, \mathbf{q}^T|\tilde{\Theta}) \tag{15}$$

donde $\Theta$ es la estimación inicial que se posee para el modelo y $\tilde{\Theta}$ es la nueva estimación. La normalización mediante $\mathbf{X}^T$ permite aplicar la función auxiliar a múltiples secuencias de entrenamiento (como se verá en la Sección 2.7).

El algoritmo de *maximización de la esperanza* es un caso particular del método de máxima verosimilitud que posee menor costo computacional [Duda et al., 1999]. Este algoritmo se basa en iterar haciendo en cada paso $\tilde{\Theta}$ igual a aquel $\Theta$ que haya maximizado la función auxiliar $\mathcal{O}$ en el paso anterior. Como requisito de convergencia, si en cualquier paso del algoritmo se verifica $\mathcal{O}(\Theta, \tilde{\Theta}) \geq \mathcal{O}(\Theta, \Theta)$, entonces debe cumplirse que $\Pr(\mathbf{X}^T|\tilde{\Theta}) \geq \Pr(\mathbf{X}^T|\Theta)$. Para el caso de la función auxiliar seleccionada en (15), puede encontrarse en [Huang et al., 1990] una demostración sencilla de que esta propiedad se cumple.

Para aplicar este algoritmo a la estimación de los parámetros del MOMC se debe obtener primero la ecuación completa para $\mathcal{O}(\Theta, \tilde{\Theta})$. A partir de (9) y (11) se puede escribir:

$$p\left(\mathbf{X}^T, \mathbf{q}^T|\Theta\right) = \sum_{k_1=1}^{N_c} \sum_{k_2=1}^{N_c} \cdots \sum_{k_T=1}^{N_c} \left\{ \prod_{t=1}^{T} b_{q_t k_t}(\mathbf{x}_t)\, a_{q_{t-1} q_t} \right\} c_{q_1 k_1} c_{q_2 k_2} \cdots c_{q_T k_T}$$

y así es posible redefinir (11) como:

$$\Pr\left(\mathbf{X}^T|\Theta\right) = \sum_{\forall \mathbf{q}^T} \sum_{\forall \mathbf{c}^T} \prod_{t=1}^{T} b_{q_t k_t}(\mathbf{x}_t)\, a_{q_{t-1} q_t} c_{q_t k_t} = \sum_{\forall \mathbf{q}^T} \sum_{\forall \mathbf{c}^T} p\left(\mathbf{X}^T, \mathbf{q}^T, \mathbf{c}^T|\Theta\right)$$

donde las $\mathbf{c}^T$ son las secuencias de la forma $c_{q_1 k_1}, c_{q_2 k_2}, \ldots, c_{q_T k_T}$.

Para poder desarrollar completamente la función auxiliar de (15) queda por obtener:

[2] Número de Kullback-Leibler.

$$\log p\left(\mathbf{X}^T, \mathbf{q}^T, \mathbf{c}^T \left| \tilde{\Theta} \right.\right) = \sum_{t=1}^{T} \log \tilde{b}_{q_t k_t}\left(\mathbf{x}_t\right) + \sum_{t=1}^{T} \log \tilde{a}_{q_{t-1} q_t} + \sum_{t=1}^{T} \log \tilde{c}_{q_t k_t}$$

y así la expresión de la función auxiliar queda convenientemente separada en:

$$\begin{aligned}
\mathcal{O}(\Theta, \tilde{\Theta}) &= \frac{1}{p(\mathbf{X}^T|\Theta)} \sum_{\forall \mathbf{q}^T} \sum_{\forall \mathbf{c}^T} p\left(\mathbf{X}^T, \mathbf{q}^T, \mathbf{c}^T | \Theta\right) \\
&\quad \times \left\{ \sum_{t=1}^{T} \log \tilde{b}_{q_t k_t}\left(\mathbf{x}_t\right) + \sum_{t=1}^{T} \log \tilde{a}_{q_{t-1} q_t} + \sum_{t=1}^{T} \log \tilde{c}_{q_t k_t} \right\} \\
&= \mathcal{O}_b(\Theta, \tilde{b}_{jk}) + \mathcal{O}_a(\Theta, \tilde{a}_{ij}) + \mathcal{O}_c(\Theta, \tilde{c}_{jk})
\end{aligned}$$

con:

$$\mathcal{O}_b(\Theta, \tilde{b}_{jk}) = \sum_{j=1}^{|\mathcal{Q}|} \sum_{\forall \mathbf{c}^T} \sum_{t=1}^{T} p\left(q_t = j, k_t = k \left| \mathbf{X}^T, \Theta \right.\right) \log \tilde{b}_{jk}\left(\mathbf{x}_t\right) \tag{16}$$

$$\mathcal{O}_a(\Theta, \tilde{a}_{ij}) = \sum_{i=1}^{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \sum_{\forall \mathbf{c}^T} p\left(q_{t-1} = i, q_t = j, \mathbf{c}^T \left| \mathbf{X}^T, \Theta \right.\right) \log \tilde{a}_{ij} \tag{17}$$

$$\mathcal{O}_c(\Theta, \tilde{c}_{jk}) = \sum_{j=1}^{|\mathcal{Q}|} \sum_{\forall \mathbf{c}^T} \sum_{t=1}^{T} p\left(q_t = j, k_t = k \left| \mathbf{X}^T, \Theta \right.\right) \log \tilde{c}_{jk} \tag{18}$$

### 2.3.1. Probabilidades de transición

En primer lugar considérese la función auxiliar (17), con la que se obtendrá la fórmula de reestimación para los $a_{ij}$. En este caso hay que tener en cuenta que la optimización está condicionada a:

$$\sum_{j=1}^{|\mathcal{Q}|} \tilde{a}_{ij} \stackrel{\circ}{=} 1 \quad \forall i \in \mathcal{Q}$$

Es por esto que conviene utilizar los multiplicadores de Lagrange escribiendo:

$$\nabla_{\tilde{a}} \left( \mathcal{O}_a(\Theta, \tilde{a}_{ij}) - \sum_{i=1}^{|\mathcal{Q}|} \ell_i \left( \sum_{j=1}^{|\mathcal{Q}|} \tilde{a}_{ij} - 1 \right) \right) = 0$$

Reemplazando (17) en esta ecuación y haciendo las derivadas parciales con respecto a los $\tilde{a}_{ij}$ se tiene:

$$\sum_{i=1}^{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \sum_{\forall \mathbf{c}^T} \left\{ p\left(q_{t-1} = i, q_t = j, \mathbf{c}^T \left| \mathbf{X}^T, \Theta \right.\right) \frac{1}{\tilde{a}_{ij}} \right\} - \ell_i = 0 \qquad (19)$$

que puede maximizarse considerando individualmente todos los términos de la sumatoria sobre los $i$.

Es necesario obtener primero los multiplicadores de Lagrange $\ell_i$; multiplicando en ambos términos por los $\tilde{a}_{ij}$:

$$\sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \sum_{\forall \mathbf{c}^T} p\left(q_{t-1} = i, q_t = j, \mathbf{c}^T \left| \mathbf{X}^T, \Theta \right.\right) = \sum_{j=1}^{|\mathcal{Q}|} \tilde{a}_{ij} \ell_i$$

y así:

$$\begin{aligned} \ell_i &= \sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \sum_{\forall \mathbf{c}^T} p\left(q_{t-1} = i, q_t = j, \mathbf{c}^T \left| \mathbf{X}^T, \Theta \right.\right) \\ &= \sum_{t=1}^{T} \sum_{\forall \mathbf{c}^T} p\left(q_{t-1} = i, \mathbf{c}^T \left| \mathbf{X}^T, \Theta \right.\right) \end{aligned}$$

Volviendo a (19) ahora se obtiene:

$$\begin{aligned}\tilde{a}_{ij} &= \frac{\sum_{t=1}^{T}\sum_{\forall \mathbf{c}^T} p\left(q_{t-1}=i, q_t=j, \mathbf{c}^T \left| \mathbf{X}^T, \Theta\right.\right)}{\sum_{t=1}^{T}\sum_{\forall \mathbf{c}^T} p\left(q_{t-1}=i, \mathbf{c}^T \left| \mathbf{X}^T, \Theta\right.\right)} \\ &= \frac{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_{t-1}=i, q_t=j \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}}{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_{t-1}=i, \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}} \end{aligned} \tag{20}$$

### 2.3.2. Probabilidades de observación

Considérese ahora (18), para cuya optimización existe la restricción:

$$\sum_{k=1}^{N_c} \tilde{c}_{jk} = 1 \quad \forall j.$$

Este es un caso muy similar al de los $a_{ij}$ y la fórmula de reestimación se deduce a partir de:

$$\nabla_{\tilde{c}} \left( \mathcal{O}_c(\Theta, \tilde{c}_{kj}) - \sum_{j=1}^{|\mathcal{Q}|} \ell_j \left( \sum_{k=1}^{N_c} \tilde{c}_{jk} - 1 \right) \right) = 0,$$

se reemplaza aquí (18) y nuevamente se obtienen las derivadas parciales, se despejan los multiplicadores de Lagrange y la fórmula de reestimación queda:

$$\tilde{c}_{jk} = \frac{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t=j, k_t=k \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}}{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t=j \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}} \tag{21}$$

Para completar la estimación de las probabilidades de observación resta deducir la fórmula de reestimación para los $b_{jk}(\mathbf{x}_t)$, que estaban definidos en función de los vectores de medias $\boldsymbol{\mu}_{jk}$ y las matrices de covarianzas $\mathbf{U}_{jk}$. Anulando $\nabla_{\tilde{b}}\mathcal{O}_b(\Theta, \tilde{b}_{jk})$, se puede derivar primero con respecto a los $\tilde{\boldsymbol{\mu}}_{jk}$ y obtener:

$$0 = \frac{\partial \mathcal{O}_b(\Theta, \tilde{b}_{jk})}{\partial \tilde{\boldsymbol{\mu}}_{jk}} = \sum_{j=1}^{|\mathcal{Q}|} \sum_{\forall \mathbf{c}^T} \sum_{t=1}^{T} p\left(q_t = j, k_t = k \left| \mathbf{X}^T, \Theta \right.\right) \tilde{\mathbf{U}}_{jk}^{-1} (\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})$$

desde donde se despejan los $\tilde{\boldsymbol{\mu}}_{jk}$ quedando:

$$\tilde{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left| \Theta \right.\right)}{p\left(\mathbf{X}^T \left| \Theta \right.\right)} \mathbf{x}_t}{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left| \Theta \right.\right)}{p\left(\mathbf{X}^T \left| \Theta \right.\right)}} \tag{22}$$

De forma similar, a partir de $\nabla_{\tilde{b}}\mathcal{O}_b(\Theta, \tilde{b}_{jk}) = 0$ y derivando con respecto a los $\tilde{\mathbf{U}}_{jk}^{-1}$:

$$\begin{aligned} 0 &= \frac{\partial \mathcal{O}_b(\Theta, \tilde{b}_{jk})}{\partial \tilde{\mathbf{U}}_{jk}^{-1}} = \\ &= \sum_{j=1}^{|\mathcal{Q}|} \sum_{\forall \mathbf{c}^T} \sum_{t=1}^{T} p\left(q_t = j, k_t = k \left| \mathbf{X}^T, \Theta \right.\right) \frac{1}{2} \tilde{\mathbf{U}}_{jk}^{-1} - (\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})^{\mathrm{T}} \end{aligned}$$

de donde se despeja:

$$\tilde{\mathbf{U}}_{jk}^{-1} = \frac{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left| \Theta \right.\right)}{p\left(\mathbf{X}^T \left| \Theta \right.\right)} (\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})^{\mathrm{T}}}{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left| \Theta \right.\right)}{p\left(\mathbf{X}^T \left| \Theta \right.\right)}} \tag{23}$$

### 2.3.3. Interpretaciones de las fórmulas de reestimación

Para llegar a una interpretación conceptual de estas fórmulas de reestimación es útil definir:

I) La variable $\alpha$:

$$\alpha_t(i) \triangleq \Pr\left(\mathbf{x}_1, \ldots, \mathbf{x}_t, q_t = i \,|\Theta\right) \tag{24}$$

calculable de forma inductiva a partir de $\alpha_1(i) = b_i(\mathbf{x}_1)$ mediante $\alpha_t(j) = \sum_{\forall i \in \mathcal{Q}} \alpha_{t-1}(i) a_{ij} b_j(\mathbf{x}_t)$. Así se puede reescribir (11) como $\Pr\left(\mathbf{X}^T |\Theta\right) = \sum_{\forall i \in \mathcal{Q}} \alpha_T(i)$.

II) La variable $\beta$:

$$\beta_t(i) \triangleq \Pr\left(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T, q_t = i \,|\Theta\right) \tag{25}$$

que puede calcularse por inducción comenzando con $\beta_T(i) = 1/|\mathcal{Q}|$ y haciendo $\beta_t(j) = \sum_{\forall i \in \mathcal{Q}} a_{ji} b_i(\mathbf{x}_{t+1}) \beta_{t+1}(i)$. Ahora se puede reescribir (11) como $\Pr\left(\mathbf{X}^T |\Theta\right) = \sum_{\forall i \in \mathcal{Q}} b_i(\mathbf{x}_1) \beta_1(i)$.

III) Las variables $\gamma$:

$$\gamma_t(i) \triangleq \Pr\left(q_t = i \,\middle|\, \mathbf{X}^T, \Theta\right) = \frac{\alpha_t(i)\beta_t(i)}{\sum\limits_{\forall q \in \mathcal{Q}} \alpha_T(q)} \tag{26}$$

que puede interpretarse como la cantidad de veces que el estado $i$ es visitado en el instante de tiempo $t$, para observar la secuencia de evidencias acústicas $\mathbf{X}^T$. Además, se puede definir:

$$\begin{aligned} \gamma_t(i,j) &\triangleq \Pr\left(q_{t-1} = i, q_t = j \,\middle|\, \mathbf{X}^T, \Theta\right) \\ &= \frac{\alpha_{t-1}(i) a_{ij} \sum\limits_{k=1}^{N_c} c_{jk} b_{jk}\left(\mathbf{x}_t\right) \beta_t(j)}{\sum\limits_{\forall q \in \mathcal{Q}} \alpha_T(q)} \end{aligned} \tag{27}$$

equivalente a pensar en la cantidad de veces que se ha llegado al estado $j$ a partir del $i$, bajo las mismas condiciones anteriores.

IV) La variable $\psi$:

$$\begin{aligned} \psi_t(j,k) &\triangleq \Pr\left(q_t = j, k_t = k \,\middle|\, \mathbf{X}^T, \Theta\right) \\ &= \frac{\sum_{i\in\mathcal{Q}} \alpha_{t-1}(i) a_{ij} c_{jk} b_{jk}\left(\mathbf{x}_t\right) \beta_t(j)}{\sum_{\forall q\in\mathcal{Q}} \alpha_T(q)} \end{aligned} \tag{28}$$

interpretable como la cantidad esperada de veces en que se llegó al estado $j$ en el tiempo $t$ utilizando la gaussiana $k$, cuando se entrenaba el modelo $\Theta$ con la secuencia de evidencias acústicas $\mathbf{X}^T$.

Mediante estas definiciones pueden reescribirse las ecuaciones (20), (21), (22) y (23), respectivamente, como:

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} \qquad \tilde{c}_{jk} = \frac{\sum_{t=1}^{T} \psi_t(j,k)}{\sum_{t=1}^{T} \gamma_t(i)}$$

$$\tilde{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} \psi_t(j,k)\mathbf{x}_t}{\sum_{t=1}^{T} \psi_t(j,k)} \qquad \tilde{\mathbf{U}}_{jk}^{-1} = \frac{\sum_{t=1}^{T} \psi_t(j,k)(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})^{\mathrm{T}}}{\sum_{t=1}^{T} \psi_t(j,k)}$$

Escritas en esta forma, las fórmulas de reestimación se conocieron originalmente como parte del algoritmo de reestimación de *Baum-Welch*. En el trabajo original las probabilidades de observación eran discretas, con lo que se simplifican los $\tilde{c}_{jk}$, $\tilde{\boldsymbol{\mu}}_{jk}$ y $\tilde{\mathbf{U}}_{jk}^{-1}$ en un $\tilde{b}_j(x_k) = \sum_t \gamma_t(i)\delta(x_k, o_t) / \sum_t \gamma_t(i)$.

Por otro lado, si se realiza la búsqueda de la secuencia más probable $\tilde{\mathbf{q}}^T$ mediante el algoritmo de Viterbi (14) y se redefinen (26) y (27) de forma que solamente tomen valores 0 o 1 ($\gamma_t(i) = 1$ cuando $\tilde{q}_t = i$ y $\gamma_t(i,j) = 1$

cuando $\tilde{q}_{t-1} = i \wedge \tilde{q}_t = j$), entonces al aplicar las fórmulas de reestimación y buscar la secuencia más probable sucesivamente se obtiene el denominado algoritmo de entrenamiento de Viterbi, que posee un costo computacional mucho menor al de Baum-Welch y tiene buen rendimiento en las aplicaciones prácticas de RAH.

### 2.3.4. Extensiones para modelos semicontinuos

Los MOM semicontinuos (MOMSC) surgen para reducir el número total de parámetros a estimar durante el entrenamiento. En los MOMC las probabilidades de observación $b_{jk}(\cdot)$ podían estar representadas arbitrariamente por cualquier distribución $\mathcal{N}(\cdot)$. Ahora, los MOMSC, podrán compartir un conjunto fijo de gaussianas conservando para cada estado la posibilidad de asignar diferentes pesos $c_{jk}$ en la mezcla. Esto es conocido también como enlazado de parámetros. Se redefine (9) simplificando la dependencia entre los parámetros de $\mathcal{N}(\cdot)$ y el estado $j$:

$$b_j\left(\mathbf{x}_t\right) = \sum_{k=1}^{N_c} c_{jk} b_k\left(\mathbf{x}_t\right)$$

siendo en este caso:

$$b_k\left(\mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_t, \boldsymbol{\mu}_k, \mathbf{U}_k\right) = \frac{1}{(2\pi)^{N_x}\left|\mathbf{U}_k\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(\mathbf{x}_t-\boldsymbol{\mu}_k)^{\mathrm{T}}\mathbf{U}_k^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_k)\right]}$$

La función auxiliar para la optimización (16) ahora se simplifica y queda:

$$\mathcal{O}_b(\Theta, \tilde{b}_k) = \sum_{\forall \mathbf{c}^T} \sum_{t=1}^{T} p\left(k_t = k \left| \mathbf{X}^T, \Theta\right.\right) \log \tilde{b}_k\left(\mathbf{x}_t\right)$$

y al igual que antes, derivando e igualando a cero, se obtienen:

$$\tilde{\boldsymbol{\mu}}_k = \frac{\sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)} \mathbf{x}_t}{\sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}} \tag{29}$$

$$\tilde{\mathbf{U}}_k^{-1} = \frac{\sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \,|\Theta\right)}{p\left(\mathbf{X}^T \,|\Theta\right)} (\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_k)(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_k)^{\mathrm{T}}}{\sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \,|\Theta\right)}{p\left(\mathbf{X}^T \,|\Theta\right)}} \tag{30}$$

que en comparación con (23) y (22) simplemente se han incorporado las sumatorias sobre $j$, calculando así la probabilidad sobre todos los estados en $p\left(\mathbf{X}^T, q_t = j, k_t = k \,|\Theta\right)$.

## 2.4. Concatenación de modelos

A partir del modelo genérico $\Theta$ es posible construir un conjunto con modelos de fonemas para el RAH:

$$\mathcal{F}_\Theta = \left\{ {}^F\Theta_\varphi \right\}; \quad \varphi \in \mathcal{F}$$

donde $\mathcal{F} = [1 \ldots |\mathcal{F}_\Theta|]$ es el conjunto de los fonemas para el reconocimiento. Un modelo de palabra se define como la concatenación de varios modelos de fonemas. El último estado de cada fonema se une directamente —con probabilidad de transición 1— al primero del siguiente conformando palabras:

$${}^W\Theta_w = {}^F\Theta_{\varphi_1} {}^F\Theta_{\varphi_2} \ldots {}^F\Theta_{\varphi_{N_w}}; \quad \varphi_f \in \mathcal{F} \tag{31}$$

a partir de un diccionario de pronunciaciones o transcripciones fonéticas:

$$\mathcal{W}_\varphi = \{(w; \varphi_1, \varphi_2, \ldots, \varphi_{N_w})\}; \quad N_w < \infty; \quad w \in \mathcal{W}$$

donde $\mathcal{W} = [1 \ldots |\mathcal{W}_\varphi|]$ es el conjunto de palabras para el reconocimiento. Estos modelos compuestos (MC) pueden ser vistos como un MOM de más estados y son tratados formalmente como se describió antes. Si el conjunto de estados de un MOM se pueden obtener mediante el funcional $\mathcal{Q}(\Theta)$, la cantidad de estados de un modelo de palabra es:

$$\left|\mathcal{Q}\left({}^{W}\Theta_{w}\right)\right| = \sum_{f=1}^{N_w} \left|\mathcal{Q}\left({}^{F}\Theta_{\varphi_f}\right)\right| \tag{32}$$

Ahora se puede construir el conjunto de modelos del vocabulario de reconocimiento:

$$\mathcal{W}_{\Theta} = \left\{{}^{W}\Theta_{w}\right\}; \quad w \in \mathcal{W}$$

## 2.5. Modelado estadístico del lenguaje

Sean $M, N \in \mathbb{N}; M, N < \infty$ y sea:

$$\mathbf{w}^{M} = w_1, w_2, \ldots, w_M; \quad w_m \in \mathcal{W} \tag{33}$$

una secuencia ordenada de $M$ palabras a reconocer. Para cada palabra $w_m$ en la secuencia, se define su historia de orden $N$ como:

$$\mathbf{h}_m^N = w_{m-1}, w_{m-2}, \ldots, w_{m-N+1}; \quad w_{m-j} \in \mathcal{W}.$$

El modelo de lenguaje (ML) puede ser aproximado mediante la utilización de las denominadas $n$-gramáticas:

$$\Pr\left(\mathbf{w}^{M}\right) = \prod_{m=1}^{M} \Pr\left(w_m \left|\mathbf{h}_m^m\right.\right) \approx \prod_{m=1}^{M} \Pr\left(w_m \left|\mathbf{h}_m^N\right.\right) \triangleq G^N\left(\mathbf{w}^{M}\right) \tag{34}$$

La probabilidad de una palabra $w_m$, dada su historia $\mathbf{h}_m^N$, puede ser estimada simplemente mediante sus frecuencias de ocurrencia:

$$\Pr\left(w_m \left|\mathbf{h}_m^N\right.\right) \approx \frac{\mathcal{C}\left(w_m, \mathbf{h}_m^N\right)}{\mathcal{C}\left(\mathbf{h}_m^N\right)}$$

donde $\mathcal{C}(\cdot)$ es una función que cuenta las ocurrencias de una determinada secuencia de palabras en el corpus de entrenamiento.

Sin embargo, en muchos casos prácticos algunas historias $\mathbf{h}_m^N$ nunca aparecen en el corpus de entrenamiento. Es por esto que resulta necesario considerar el *suavizado* de las gramáticas. Por medio de estas técnicas, es posible estimar las probabilidades de las palabras cuyas historias de orden $N$ nunca aparecen en el corpus de entrenamiento. Existen muchas técnicas útiles para el suavizado de gramáticas [Jelinek, 1999]. Un primer método sencillo es el denominado suavizado por interpolación lineal [Rabiner y Juang, 1993]. Dado un $K \in \mathbb{N}, 0 \leq K \leq N-1$ y la historia:

$$\mathbf{h}_m^K \ / \ \mathcal{C}\left(\mathbf{h}_m^K\right) > 0$$

se estiman las probabilidades para las historias inexistentes mediante:

$$I_m^K = \sum_{k=0}^{K} \iota_k \Pr\left(w_m \left| \mathbf{h}_m^k \right.\right) \tag{35}$$

con $0 \leq \iota_k \leq 1$ y $\sum \iota_k = 1$. Las historias $\mathbf{h}^1$ corresponden a una unigramática y la probabilidad para el caso de las historias $\mathbf{h}^0$ se define como:

$$\Pr\left(w_m \left| \mathbf{h}_m^0 \right.\right) \triangleq \frac{1}{|\mathcal{W}|} \quad \forall w_m \in \mathcal{W}.$$

Una de las técnicas más utilizadas para la estimación y suavizado de gramáticas es la denominada *back-off* [Potamianos y Jelinek, 1998]:

$$\Omega_m^K = \begin{cases} \dfrac{\mathcal{C}(w_m, \mathbf{h}_m^K) - \vartheta}{\mathcal{C}(\mathbf{h}_m^K)} & \text{si } \mathcal{C}\left(w_m, \mathbf{h}_m^K\right) > 0 \\ \varsigma(\mathbf{h}_m^K)\Omega_m^{K-1} & \text{si } \mathcal{C}\left(w_m, \mathbf{h}_m^K\right) = 0 \end{cases} \tag{36}$$

donde se fija empíricamente $\vartheta = 0{,}5$.

Para encontrar las probabilidades $\varsigma(\mathbf{h}_m^K)$ se debe considerar primeramente que:

$$\sum_{w_m / \mathcal{C}(w_m, \mathbf{h}_m^K) > 0} \frac{\mathcal{C}\left(w_m, \mathbf{h}_m^K\right) - \vartheta}{\mathcal{C}\left(\mathbf{h}_m^K\right)} + \sum_{w_m / \mathcal{C}(w_m, \mathbf{h}_m^K) = 0} \varsigma\left(\mathbf{h}_m^K\right) \Omega_m^{K-1} = 1$$

de esta forma:

$$\varsigma\left(\mathbf{h}_m^K\right)\left(\sum_{w_m/\mathcal{C}(w_m,\mathbf{h}_m^K)=0}\Omega_m^{K-1}\right)=\left(1-\sum_{w_m/\mathcal{C}(w_m,\mathbf{h}_m^K)>0}\frac{\mathcal{C}\left(w_m,\mathbf{h}_m^K\right)-\vartheta}{\mathcal{C}\left(\mathbf{h}_m^K\right)}\right)$$

y así:

$$\varsigma\left(\mathbf{h}_m^K\right)=\frac{1-\sum_{w_m/\mathcal{C}(w_m,\mathbf{h}_m^K)>0}\Omega_m^K}{1-\sum_{w_m/\mathcal{C}(w_m,\mathbf{h}_m^K)>0}\Omega_m^{K-1}}$$

## 2.6. Decodificación en el modelo compuesto

El MC es una estructura en red con todos los modelos de palabra conectados a partir de las probabilidades del ML. También es posible ver al MC como un gran MOM; si $|\mathcal{Q}|_{(m)}$ es el último estado del modelo de palabra ${}^{W}\Theta_{w_m}$ y $1_{(n)}$ el primero de ${}^{W}\Theta_{w_n}$, entonces se define la probabilidad de transición entre las dos palabras del MC como:

$$a_{|\mathcal{Q}|_{(m)},1_{(n)}}\triangleq G_{mn}^{(2)} \tag{37}$$

quedando así definida la estructura del MC ${}^{C}\Theta$ para una frase completa[3] o, si se quiere, para cualquier frase posible dado el conjunto de palabras $\mathcal{W}$ y el ML que la relaciona.

En la extensión del algoritmo de Viterbi se requiere incorporar las probabilidades del ML en el proceso de búsqueda sobre el MC. Dadas las palabras $w_m, w_n \in \mathcal{W}$, se utilizará la siguiente notación:

$i_{(m)}, j_{(m)}$: estados pertenecientes al modelo de la palabra $w_m$,

$q_{(m)t}$: estado de ${}^{W}\Theta_{w_m}$ en el tiempo $t$,

[3]La diferencia de esta concatenación de modelos en relación a (31) radica en que la probabilidad de transición entre dos modelos de palabra queda definida por el ML mientras que la probabilidad de transición entre fonemas era siempre 1.

$\mathbf{q}_{(m)}^{T}$: secuencia de $T$ estados en ${}^{W}\Theta_{w_m}$ y

$G_{mn}^{(2)}$: probabilidad de que se emita $w_n$ con una historia $\mathbf{h}_n^2 = w_m$ (ver ecuación (34))

Considerando un ML de bi-gramática es posible redefinir la probabilidad acumulada de la ecuación (12) como:

$$\Lambda_t(j_{(n)}) \triangleq \max_{\forall \mathbf{q}_{(n)}^{t-1}} \left\{ \Pr\left( \mathbf{q}_{(n)}^{t}, q_{(n)t} = j_{(n)}, \mathbf{X}^t \left| {}^{W}\Theta_{w_m} \right. \right) \right\}$$

con las inicializaciones:

$$\Lambda_0(j_{(n)}) = 1 \quad \forall w_n \in \mathcal{W}, \forall j_{(n)} \in \mathcal{Q}\left({}^{W}\Theta_{w_n}\right)$$

y cuando comienza cada palabra[4]:

$$\Lambda_{t-1}(j_{(n)} = 1) = \max_{\forall w_m \in \mathcal{W}} \left\{ \Lambda_{t-1}\left( i_{(m)} = \left| \mathcal{Q}\left({}^{W}\Theta_{w_m}\right) \right| \right) G_{mn}^{(2)} \right\}.$$

Luego, es posible expandir esta probabilidad acumulada como:

$$\Lambda_t\left(j_{(n)}\right) = \max_{\forall \mathbf{q}_{(n)}^{t-1}} \left\{ \Pr\left( \mathbf{q}_{(n)}^{t-1}, \mathbf{x}^t \left| {}^{W}\Theta_{w_n} \right. \right) \Pr\left( q_{(n)t} = j_{(n)}, \mathbf{X}^t \left| \mathbf{q}_{(n)}^{t-1}, {}^{W}\Theta_{w_n} \right. \right) \right\}$$

y calcularla por inducción mediante:

$$\begin{aligned}
\Lambda_t\left(j_{(n)}\right) &= \max_{\forall i_{(n)}} \left\{ \Lambda_{t-1}(i_{(n)}) \Pr\left( q_{(n)t} = j_{(n)}, \mathbf{x}_t \left| q_{(n)t-1} = i_{(n)}, {}^{W}\Theta_{w_n} \right. \right) \right\} \\
&= \max_{\forall i_{(n)}} \left\{ \Lambda_{t-1}(i_{(n)}) \Pr\left( q_{(n)t} = j_{(n)} \left| q_{(n)t-1} = i_{(n)}, {}^{W}\Theta_{w_n} \right. \right) \right. \\
&\qquad \left. \times \Pr\left( \mathbf{x}_t \left| q_{(n)t} = j_{(n)}, {}^{W}\Theta_{w_n} \right. \right) \right\} \\
&= \max_{\forall i_{(n)}} \left\{ \Lambda_{t-1}(i_{(n)}) \Pr\left( q_{(n)t} = j_{(n)} \left| q_{(n)t-1} = i_{(n)}, {}^{W}\Theta_{w_n} \right. \right) \right\} \\
&\quad \times \Pr\left( \mathbf{x}_t \left| q_{(n)t} = j_{(n)}, {}^{W}\Theta_{w_n} \right. \right)
\end{aligned}$$

[4]Obsérvese que en la transición entre dos palabras el modelo no emite y por lo tanto no cambia el índice de tiempo $t$.

$$\Lambda_t\left(j_{(n)}\right) \quad = \quad \underset{\forall i_{(n)}}{\text{máx}} \left\{ \Lambda_{t-1}(i_{(n)}) a_{i_{(n)} j_{(n)}} \right\} b_{j_{(n)}}\left(\mathbf{x}_t\right)$$

Para obtener la secuencia más probable a partir de las probabilidades acumuladas se define:

$$\Xi_t(j_{(n)} \triangleq \arg \underset{\forall i_{(n)}}{\text{máx}} \left\{ \Lambda_{t-1}(i_{(n)}) a_{i_{(n)} j_{(n)}} \right\}$$

con la salvedad de que:

$$\Xi_t\left(j_{(n)} = \left|\mathcal{Q}\left({}^{W}\Theta_{w_n}\right)\right|\right) = \arg \underset{\forall i_{(m)}=1}{\text{máx}} \left\{ \Lambda_t(i_{(m)}) G^{(2)}_{mn} \right\}$$

Ahora, por recursión inversa:

$$\tilde{q}_{(n)t} = \Xi_{t+1}\left(\tilde{q}_{(n)t+1}\right); \quad t = T-1, T-2, \ldots, 1$$

comenzando por:

$$\tilde{q}_{(n)T} = \arg \underset{\forall i_{(n)}, \forall w_n}{\text{máx}} \left\{ \Lambda_T(i_{(n)}) \right\}.$$

y con las restricciones:

$$\tilde{q}_{(n)T} \stackrel{\circ}{=} |\mathcal{Q}|_{(n)} \qquad \wedge \qquad \tilde{q}_{(n)1} \stackrel{\circ}{=} 1_{(n)}$$

La secuencia resultante está restringida por este algoritmo a una secuencia de palabras válidas ya que los fonemas están concatenados en palabras (31) y no hay conexiones hacia afuera de las palabras que no sean a través de las conexiones impuestas por el ML (siempre desde el último estado de una palabra hacia el primero de otra). Por lo tanto, dado que en esta secuencia quedan especificados tanto el número de estado como la palabra

a la que cada uno pertenece, se puede extraer directamente de ella la transcripción reconocida. Estas ecuaciones son la base del denominado algoritmo de *decodificación* para RAH. Se agregan además mejoras de índole práctico como el escalado o la aritmética logarítmica para reducir los errores introducidos por la precisión limitada en el cómputo [Rabiner y Juang, 1993]. Otras mejoras ampliamente utilizadas son las técnicas de podado, que reducen significativamente el espacio de la búsqueda en el algoritmo de Viterbi. Por ejemplo, en el algoritmo de *beam search* se utiliza una probabilidad $\Phi_\Lambda$ como umbral de podado y no se consideran los caminos que acumulan una probabilidad $\Phi_\Lambda$ veces menor que el máximo para cada tiempo $t$. Puede consultarse una revisión acerca de estos métodos en [Ney y Ortmanns, 1999].

## 2.7. Entrenamiento del modelo compuesto

Es necesario dar respuesta a tres cuestiones importantes para encontrar las fórmulas de reestimación en el MC. La primera tiene que ver con la relación entre el entrenamiento de los MOM de cada fonema y la estimación de las probabilidades del ML. La segunda cuestión se plantea al considerar múltiples secuencias —es decir, muchas frases— de entrenamiento, ya que las fórmulas de reestimación siempre se dedujeron a partir de una única secuencia de evidencias acústicas. La tercera cuestión tiene que ver con la forma en que los diferentes MOMSC, que forman el MC, van a compartir sus parámetros y las modificaciones que esto demanda en las fórmulas de reestimación.

La solución práctica más empleada para la primera cuestión es muy simple y consiste en estimar las probabilidades asociadas con el ML separadamente (por ejemplo mediante (35) o (36)), dejándolas fijas durante las reestimaciones de todos los restantes parámetros del MC [Young et al., 2000].

Para extender las fórmulas de reestimación a múltiples secuencias de evidencias acústicas, considérese que existen $N_X$ secuencias de entrenamiento:

$$\mathbf{X} = \mathbf{X}_1^{T_1}, \mathbf{X}_2^{T_2}, \ldots, \mathbf{X}_{N_X}^{T_{N_X}}$$

Bajo la hipótesis de independencia estadística entre las diferentes secuencias, la ecuación (10) debe reescribirse como:

$$\Pr\left(\mathbf{X}\,|\Theta\right) = \prod_{n=1}^{N_X} \sum_{\forall \mathbf{q}_n^{T_n}} \Pr\left(\mathbf{X}_n^{T_n}, \mathbf{q}_n^{T_n}\,|\Theta\right)$$

lo cual agrega simplemente una sumatoria sobre todas las secuencias tanto en el numerador como en el denominador de las fórmulas de reestimación. Por ejemplo, para las probabilidades de transición:

$$\tilde{a}_{ij} = \frac{\sum_{n=1}^{N_X} \sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t-1} = i, q_{n,t} = j\,|\Theta\right)}{p\left(\mathbf{X}_n^{T_n}\,|\Theta\right)}}{\sum_{n=1}^{N_X} \sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t-1} = i,|\Theta\right)}{p\left(\mathbf{X}_n^{T_n}\,|\Theta\right)}} \tag{38}$$

Durante el proceso de entrenamiento, además de contar con las secuencias de evidencias acústicas $\mathbf{X}$, también se poseen las transcripciones en palabras para cada secuencia:

$$\mathbf{W} = \mathbf{w}_1^{T_1}, \mathbf{w}_2^{T_2}, \ldots, \mathbf{w}_{N_X}^{T_{N_X}}$$

donde cada transcripción $\mathbf{w}_n^{T_n}$ es una secuencia de $T_n$ palabras como en (33):

$$\mathbf{w}_n^{T_n} = w_{n,1}, w_{n,2}, \ldots, w_{n,T_n}; \quad w_{n,m} \in \mathcal{W}$$

A partir de una de estas transcripciones y del diccionario fonético $\mathcal{W}_\varphi$ es posible construir un MC con la concatenación de palabras:

$${}^{C}\Theta_n = {}^{W}\Theta_{w_{n,1}}\,{}^{W}\Theta_{w_{n,2}} \ldots {}^{F}\Theta_{w_{n,T_n}}$$

con probabilidades fijas entre las palabras. De forma similar a (37), se puede hacer:

$$a_{|\mathcal{Q}|_{(m-1)},1_{(m)}} \triangleq P$$

donde $P$, en general, es 1.

A partir de cada uno de los MC construidos, deben estimarse todos los parámetros de los MOM que los componen. En este esquema de entrenamiento debe considerarse que el mismo modelo de fonema o palabra aparecerá en distintas partes del MC y en distintos MC para distintas frases. Al considerar que en uno de estos MC existen conjuntos de estados que comparten sus parámetros surge naturalmente la tercera cuestión, acerca de las diversas formas de compartir los parámetros en el MC. Se podrían compartir los parámetros correspondientes a los estados de una misma palabra o de un mismo fonema. También se podrían compartir parámetros de sonidos similares desde el punto de vista de la fonética acústica o bien utilizar métodos automáticos para encontrar qué conjunto de estados conviene que compartan parámetros.

A continuación se va a considerar que los estados que comparten parámetros se agrupan en conjuntos $\mathcal{Q}_{(m)}$. Estos conjuntos de estados se encontrarán previamente definidos según algún criterio y se utilizará una extensión de la notación $i_{(m)}$ y $j_{(m)}$ para indicar que estos estados pertenecen a la clase $m$. Anteriormente, se utilizaron subíndices similares para indicar la pertenencia de un estado al conjunto de estados de una palabra. Ahora, en un sentido más amplio, una clase $m$ puede corresponderse con cualquier conjunto de estados arbitrariamente agrupados[5]. De forma similar, como cada clase $m$ posee su propia mezcla de gaussianas, se deben definir los conjuntos de mezclas de gaussianas $\mathcal{M}_{(m)}$, cada uno con $N_{c_{(m)}}$ gaussianas[6]. Para indicar la pertenencia de una gaussiana $k$ al conjunto de gaussianas de la clase $m$ se utilizará la notación $k_{(m)}$.

Así como en (29) y (30) se compartían los parámetros de las mezclas de gaussianas entre los estados de un único modelo, ahora se generaliza la idea de MOMSC hacia los MC con múltiples secuencias. Siguiendo de (20) y (38), las probabilidades de transición entre los estados $i_{(m)}, j_{(m)} \in \mathcal{Q}_{(m)}$, se reestiman mediante:

$$\tilde{a}_{i_{(m)}j_{(m)}} = \frac{\sum_{n=1}^{N_X} \sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t-1} = i_{(m)}, q_{n,t} = j_{(m)} \,\middle|^C \Theta_n\right)}{p\left(\mathbf{X}_n^{T_n} \,\middle|^C \Theta_n\right)}}{\sum_{n=1}^{N_X} \sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t-1} = i_{(m)}, \,\middle|^C \Theta_n\right)}{p\left(\mathbf{X}_n^{T_n} \,\middle|^C \Theta_n\right)}} \tag{39}$$

[5]Las palabras como entidades independientes han desaparecido en los MC para el entrenamiento, salvo la situación particular en que las clases $m$ coincidan con las palabras, para lo cual la notación tampoco es contradictoria.

[6]En general $N_{c_{(m)}}$ es el mismo para todas las clases.

En el caso del peso de la gaussiana $k_{(m)}$ con que se modela la probabilidad de observación del estado $j_{(m)}$, a partir de (21):

$$\tilde{c}_{j_{(m)}k_{(m)}} = \frac{\sum_{n=1}^{N_X}\sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t}=j_{(m)}, k_t=k_{(m)} \left|{}^{C}\Theta_n\right.\right)}{p\left(\mathbf{X}_n^{T_n}\left|{}^{C}\Theta_n\right.\right)}}{\sum_{n=1}^{N_X}\sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t}=j_{(m)} \left|{}^{C}\Theta_n\right.\right)}{p\left(\mathbf{X}_n^{T_n}\left|{}^{C}\Theta_n\right.\right)}} \tag{40}$$

Al igual que en (28), se pueden simplificar las expresiones definiendo:

$$\psi_{n,t}(j_{(m)}, k_{(m)}) = \Pr\left(q_{n,t}=j_{(m)}, k_{n,t}=k_{(m)} \left|\mathbf{X}_n^{T_n}, {}^{C}\Theta_n\right.\right)$$

Dado que los parámetros de las gaussianas se comparten para una misma clase $m$, a partir de (29):

$$\tilde{\boldsymbol{\mu}}_{k_{(m)}} = \frac{\sum_{n=1}^{N_X}\sum_{\forall j_{(m)}\in\mathcal{Q}_{(m)}({}^{C}\Theta_n)}\sum_{t=1}^{T_n}\psi_{n,t}(j_{(m)},k_{(m)})\mathbf{x}_{n,t}}{\sum_{n=1}^{N_X}\sum_{\forall j_{(m)}\in\mathcal{Q}_{(m)}({}^{C}\Theta_n)}\sum_{t=1}^{T_n}\psi_{n,t}(j_{(m)},k_{(m)})} \tag{41}$$

y a partir de (30):

$$\tilde{\mathbf{U}}_{k_{(m)}}^{-1} =$$

$$= \frac{\sum_{n=1}^{N_X}\sum_{\forall j_{(m)}\in\mathcal{Q}_{(m)}({}^{C}\Theta_n)}\sum_{t=1}^{T_n}\psi_{n,t}(j_{(m)},k_{(m)})(\mathbf{x}_{n,t}-\tilde{\boldsymbol{\mu}}_{j_{(m)}k_{(m)}})(\mathbf{x}_{n,t}-\tilde{\boldsymbol{\mu}}_{j_{(m)}k_{(m)}})^{\mathrm{T}}}{\sum_{n=1}^{N_X}\sum_{\forall j_{(m)}\in\mathcal{Q}_{(m)}({}^{C}\Theta_n)}\sum_{t=1}^{T_n}\psi_{n,t}(j_{(m)},k_{(m)})} \tag{42}$$

## Referencias

[Akaike, 1974] Akaike, H. "A new look at the statistical model identification". *IEEE Trans. on Automatic Control*, volumen 19, número 6, páginas 716–723.

[Davis y Mermelstein, 1980] Davis, S. B. y Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in contunuously spoken sentences". *IEEE Trans. on Acoust. Speech, Signal Processing*, volumen 28, número 4, páginas 357–366.

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

[Duda et al., 1999] Duda, R. O., Hart, P. E., y Stork, D. G. *Pattern Classification.* John Wiley and Sons, 2° edición.

[Hess, 1991] Hess, W. J. "Pitch and voicing determination". En Furui, S. y Sondhi, M. M., editores, *Advances in Speech Signal Processing*, páginas 3–48. Marcel-Dekker, New York.

[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.

[Kay y Marple, 1981] Kay, S. M. y Marple, S. L. "Spectrum analisis". En *Proceedings of the IEEE*, volumen 69, páginas 1380–1419.

[Kuc, 1988] Kuc, R. *Introduction to digital signal processing.* McGraw-Hill Book Company.

[Makhoul, 1975] Makhoul, J. "Linear prediction: A tuturial review". En *Proceedings of the IEEE*, volumen 63, páginas 561–580.

[Ney y Ortmanns, 1999] Ney, H. y Ortmanns, S. "Dynamic programming search for continuous speech recognition". *IEEE Signal Processing Magazine*, volumen 16, número 5, páginas 64–83.

[Noll, 1967] Noll, A. M. "Cepstrum pitch determination". *Journal of the Acoustic Society of America*, volumen 41, páginas 293–309.

[Oppenheim y Schafer, 1989] Oppenheim, A. V. y Schafer, R. W. *Discrete-Time Signal Processing.* Prentice-Hall, Inc., Englewood Cliffs, NJ.

[Potamianos y Jelinek, 1998] Potamianos, G. y Jelinek, F. "A study of n-gram and decision tree letter language modeling methods". *Speech Communication*, volumen 24, páginas 171–192.

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

[Shimamura y Kobayashi, 2001] Shimamura, T. y Kobayashi, H. "Weighted autocorrelation for pitch extraction of noisy speech". *IEEE Trans. on Speech and Audio Processing*, volumen 9, número 7, páginas 727–730.

[Stevens, 1998] Stevens, K.Ñ. *Acoustic Phonetics*. MIT Press.

[Young et al., 2000] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., y Woodland, P. *HMM Toolkit*. Cambridge University, http://htk.eng.cam.ac.uk.