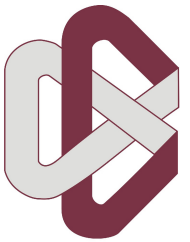
Segmentación de interlocutores a partir de señales de audio utilizando cadenas escondidas de Markov y técnicas de selección automática de modelos

Rafael de Jesús Robledo Juárez
rrobledo@cimat.mx

Asesor: Dr. Salvador Ruíz Correa



CIMAT
Centro de Investigación en Matemáticas, Guanajuato
Departamento de Ciencias de la Computación

xx de noviembre del 2013

Contenido

Contenido

Problema

- Se considera que se tiene una señal de audio con información de nuestro interés, y se requiere segmentar de acuerdo a las personas que participan en la grabación.
- *Speaker Diarization*: el problema consiste en identificar el número de interlocutores que participan en una grabación de audio, y además encontrar en qué segmentos de la grabación habla cada persona.
- Dos tareas principales:
  1. Encontrar el número total de personas que hablan en la conversación.
  2. Identificar los momentos en los que habla cada participante.

Motivación

- La tarea de speaker diarization es importante en diferentes procesos que se realizan con las grabaciones de audio, tales como la identificación y navegación por segmentos en específico.
- También resulta útil para la búsqueda y recuperación de información en grandes volúmenes de secuencias de audio.
- Es una etapa importante en el procesamiento de voz. Tanto para reconocimiento como transcripción de voz.

Principales enfoques

De acuerdo al trabajo desarrollado hasta ahora, se pueden distinguir dos grandes enfoques:

Bottom-up: Se inicia la estimación con pocos clústers (e incluso un segmento único)

Top-down: Se inicia la estimación con muchos más grupos de los que se esperan encontrar.

Ambas metodologías iteran hasta converger a un número de clústers óptimo, en que cada grupo debe corresponder a un interlocutor.

Trabajo previo

You can create overlays...

- using the `pause` command:
  - First item.
  - Second item.
- using overlay specifications:
  - First item.
  - Second item.
- using the general `uncover` command:
  - First item.
  - Second item.

Contenido

Formulación matemática (I)

Denótese por $\mathcal{A}$ la evidencia acústica a partir de la cuál el modelo deberá encontrar la segmentación correcta para un fragmento de señal.

Se puede pensar en $\mathcal{A}$ como la secuencia de símbolos correspondiente a un segmento de señal, y que está conformada por elementos de un alfabeto mucho más grande $\mathbb{A}$.

$$\mathcal{A} = a_1, a_2, ..., a_K \quad a_i \in \mathbb{A} \tag{1}$$

en donde los elementos a_i hacen referencia a un intervalo de tiempo i en la secuencia de audio original.

Formulación matemática (II)

De la misma manera,

$$\mathcal{S} = s_1, s_2, ..., s_N \quad s_i \in \mathbb{S} \tag{2}$$

donde $\mathcal{S}$ es la secuencia que corresponde a una propuesta de segmentación para un intervalo del audio original.

$\mathbb{S}$ es el conjunto de todos los interlocutores que participan en la grabación de audio y s_i de igual manera representa al interlocutor que habla en el tiempo i.

Formulación matemática (III)

Si $P(\mathcal{S} \mid \mathcal{A})$ es la probabilidad de que una secuencia de interlocutores $\mathcal{S}$ esté hablando dada la evidencia acústica en $\mathcal{A}$, entonces para escoger cuáles son los personas que hablan en ese intervalo se calcula:

$$\hat{\mathcal{S}} = \underset{\mathcal{S}}{arg\ max}\ P(\mathcal{S} \mid \mathcal{A}) \tag{3}$$

con lo que se seleccionaría la sucesión de interlocutores más probable para una secuencia de datos dada.

Notar que por el Teorema de Bayes, que (3) es equivalente a maximizar:

$$\hat{\mathcal{S}} = \underset{\mathcal{S}}{arg\ max}\ P(\mathcal{S}) \cdot P(\mathcal{A} \mid \mathcal{S}) \tag{4}$$

pues la variable $\mathcal{A}$ es constante respecto a $\mathcal{S}$.

Componentes del sistema (I)

En general, todos los sistemas que involucran procesamiento de voz, tienen varias etapas esenciales, como se menciona en Jelinek (cita)

1. Procesamiento acústico: Se refiere a la forma en la que se procesará la información y se digitalizará.

   Además se debe realizar algún proceso para obtener una representación paramétrica de la señal. A este procedimiento se le conoce como construcción del *diccionario de palabras*.

2. Modelado acústico: Se considera que ya se ha construido el diccionario de palabras o la evidencia acústica $\mathbb{A}$, y se necesita proponer una forma de calcular las probabilidades $P(\mathcal{A} \mid \mathcal{S})$.

   El modelo acústico más comúnmente utilizado en tareas de procesamiento de voz, es el HMM, aunque hay trabajos que utilizan otras técnicas: ANN [**?**] [**?**] o métodos de DTW [**?**]

Componentes del sistema (II)

3. Modelado de interlocutores: Por otro lado, se tiene que estimar también $P(\mathcal{S})$, la probabilidad de que una secuencia $\mathcal{S}$ de interlocutores participe en ese orden a priori.

   De la misma manera, por el teorema de Bayes, y puesto que deseamos calcular la probabilidad $P(\mathcal{S})$ en ese orden, se puede reescribir entonces como

$$P(\mathcal{S}) = \Pi_{i=1}^{K} P(s_i \mid s_1, ..., s_{i-1}) \tag{5}$$

   en donde se considera que el valor de s_i depende de toda la secuencia de locutores hasta ahora, situación que puede parecer algo estricta.

   Si se piensa que el interlocutor s_i sólo depende de un subconjunto más pequeño $\phi(s_1, ..., s_{i-1})$, entonces se tiene que

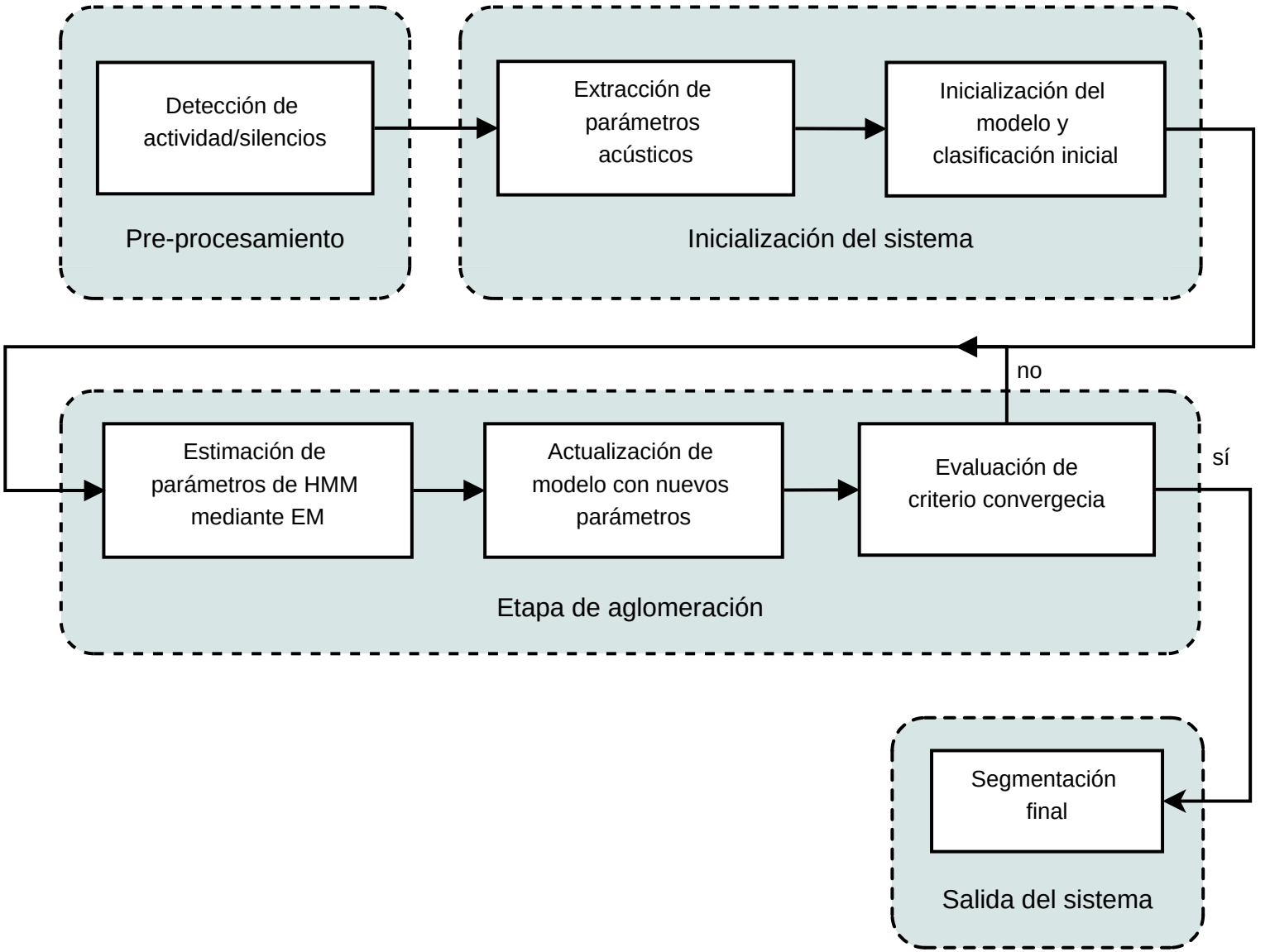
$$P(\mathcal{S}) = \Pi_{i=1}^{K} P(s_i \mid \phi(s_1, ..., s_{i-1})) \tag{6}$$

Componentes del sistema (III)

4. Búsqueda de hipótesis: En esta etapa, se deberá buscar de entre todos los posibles $\mathcal{S}$ cuál es el que maximiza (4). Como ya se mencionó, el espacio de búsqueda es realmente muy grande; por lo que no se puede hacer una búsqueda exhaustiva y deberá de seguirse una estrategia basada en la información que provee $\mathcal{A}$.

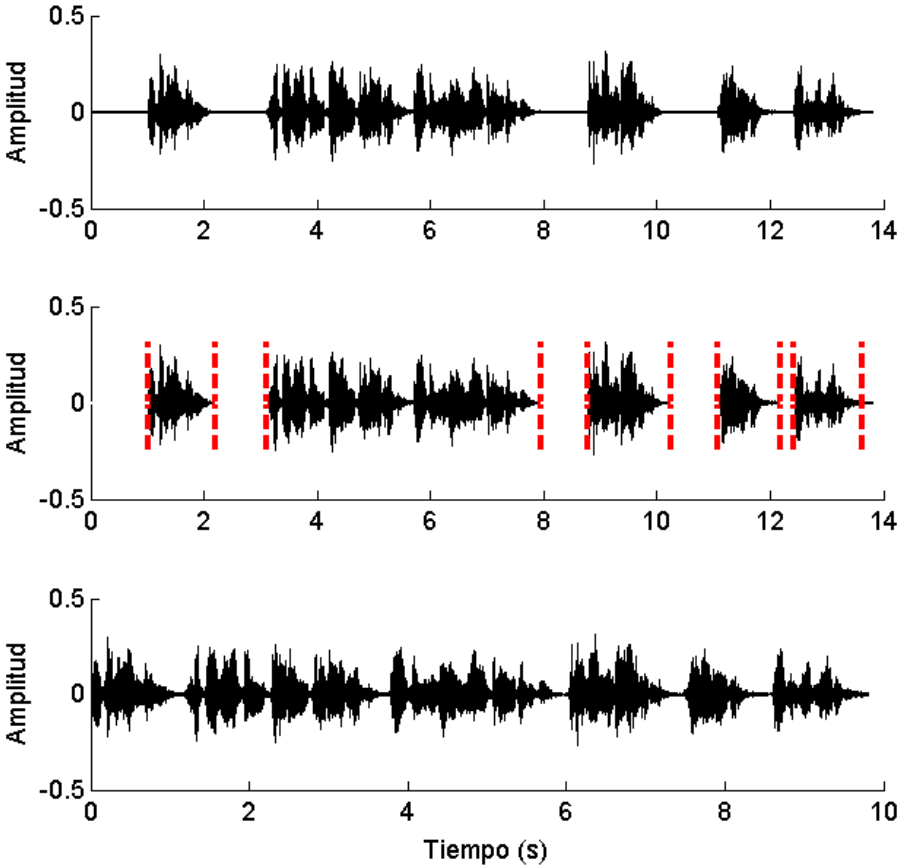
   Por otra parte, y después de obtener la segmentación correspondiente para la señal de audio, se deberá ahora inferir cuál es el modelo más probable de entre todos los que se propongan.

Componentes del sistema



Procesamiento acústico

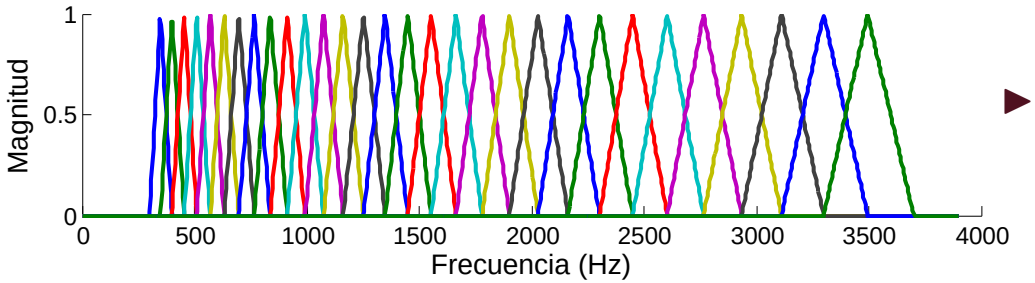
Pre-procesamiento de la señal



- Señal original
- Identificación de silencios (SAD)
- Señal de audio recortada

Procesamiento acústico

Obtención de vectores característicos (I)

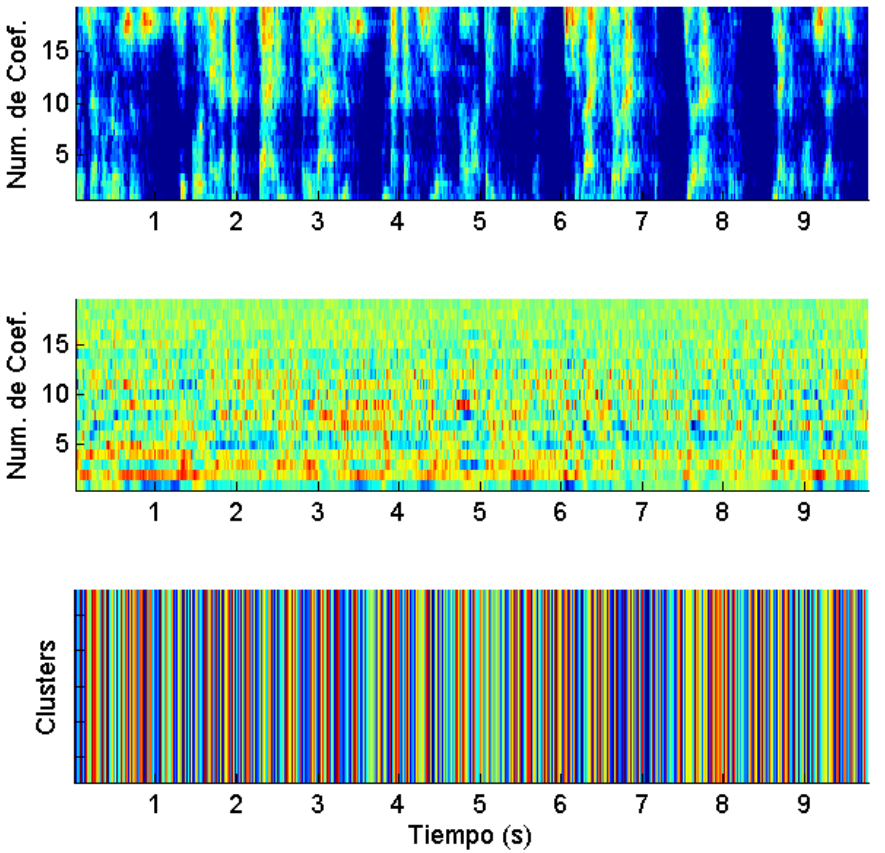


- Banco de filtros

- Se utilizará un banco de filtros (usualmente triangulares) espaciados en la escala Mel.
- Se puede calcular de la siguiente manera: $m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$
- Hay varios parámetros al momento diseñar el banco: número de filtros, frecuencia mínima y máxima, así como el ancho de banda de cada filtro.

Procesamiento acústico

Obtención de vectores característicos (II)

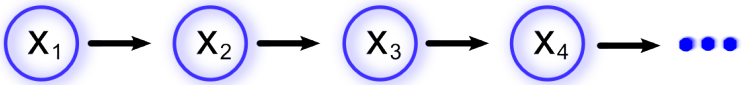


- FFT + FilterBank + log(POW)
- DCT
- k-means++

Contenido

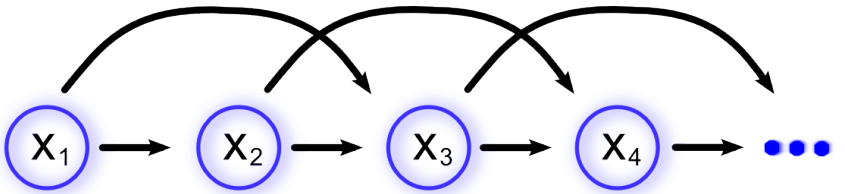
Cadenas de Markov

- Cadena de Markov de primer orden



$$p(x_1, ..., x_T) = \prod_{t=1}^{T} p(x_T \mid x_1, ..., x_{t-1}) = p(x_1) \cdot \prod_{t=2}^{T} p(x_T \mid x_{t-1}) \quad (7)$$

- Se puede generalizar para cadenas de Markov de un orden mayor



$$p(x_1, ..., x_T) = p(x_1)p(x_2 \mid x_1) \cdot \prod_{t=3}^{T} p(x_T \mid x_{t-1}, x_{t-2}) \quad (8)$$

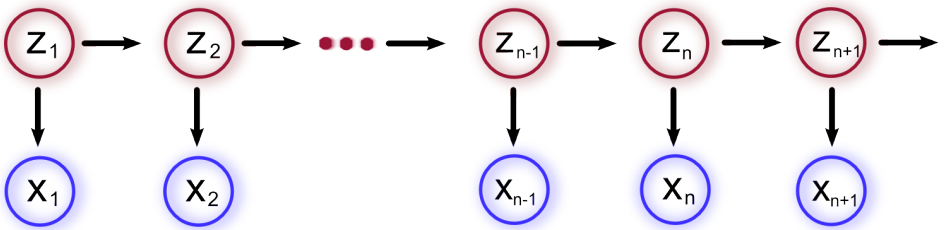
Modelo oculto de Markov

- Agregar una variable latente z_t (discreta), que corresponda a cada observación x_t.

$$z_{t+1} \perp z_{t-1} \mid z_t \tag{9}$$

$$p(x_1, ..., x_T, z_1, ..., z_T) = p(z_1) \left[\prod_{t=2}^{T} p(z_t \mid z_{t-1}) \right] \prod_{t=1}^{T} p(x_t \mid z_t). \tag{10}$$

- Modelar proceso bivariado en el tiempo. Una variable observada y una variable latente asociada.



- Mezcla de distribuciones en la que la densidad está dada por $p(x|z)$

Parámetros del HMM

- Probabilidad de cambio entre estados dada una matriz de transición **A**

$$A_{jk} \equiv p(z_{tk} = 1 \mid z_{t-1,j} = 1) \tag{11}$$

$$p(z_t \mid z_{t-1}, \mathbf{A}) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j} \cdot z_{t,k}} \tag{12}$$

- Vector de distribución inicial π para variable latente.

$$\pi_k \equiv p(z_{1k}) \tag{13}$$

$$p(z_1 \mid \pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \tag{14}$$

- Probabilidad de emisión de una variable observada x_T dada una variable latente z_T.

$$p(x_t \mid z_t, \phi) = \prod_{k=1}^{K} p(x_T \mid \phi_k)^{z_{tk}} \tag{15}$$

HMM con EM

- Probabilidad conjunta del modelo

$$p(\mathbf{X}, \mathbf{Z} \mid \theta) = p(z_1 \mid \pi) \left[\prod_{t=2}^{T} p(z_t \mid z_{t-1}, \mathbf{A}) \right] \prod_{t=1}^{T} p(x_t \mid z_t, \mathbf{B}, \phi) \qquad (16)$$

donde $\mathbf{X} = \{x_1, ..., x_N\}$, $\mathbf{Z} = \{z_1, ..., z_N\}$

y los parámetros del modelo $\theta = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}, \phi\}$

HMM con EM

- Función de verosimilitud completa

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{old}) \log p(\mathbf{X}, \mathbf{Z} \mid \theta) \tag{17}$$

Probabilidad marginal de una variable latente

$$\gamma(z_t) = p(z_t \mid \mathbf{X}, \theta^{old}) \tag{18}$$

Probabilidad conjunta de dos variables latentes consecutivas

$$\xi(z_{t-1}, z_T) = p(z_{t-1}, z_T \mid \mathbf{X}, \theta^{old}) \tag{19}$$

HMM con EM

- Prob. marginal de $z_{tk} = 1$, prob. conjunta de $z_{t-1,j}, z_{tk}$

$$\gamma(z_{tk}) = \mathbb{E}\left[z_{tk}\right] = \sum_{Z} \gamma(\mathbf{z}) z_{tk} \tag{20}$$

$$\xi(z_{t-1,j}, z_{tk}) = \mathbb{E}\left[z_{t-1,j} \cdot z_{tk}\right] = \sum_{Z} \gamma(\mathbf{z}) z_{t-1,j} \cdot z_{tk} \tag{21}$$

- Función de verosimilitud completa (reescrita con (20), (21))

$$\begin{aligned} \mathcal{Q}(\theta, \theta^{old}) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{t-1,j}, z_{tk}) \log A_{jk} + \\ \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma(z_{tk}) \log p(x_T \mid \phi_k) \end{aligned} \tag{22}$$

- Parámetros estimados por EM:

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}, \quad A_{jk} = \sum_{t=2}^{T} \frac{\xi(z_{t-1,j}, z_{tk})}{\sum_{l=1}^{K} \xi(z_{t-1,j}, z_{tl})} \tag{23}$$

Algoritmo backward-forward

$$\gamma(z_t) = p(z_t \mid X) = \frac{p(X \mid z_t)p(z_t)}{p(X)} \tag{24}$$

$$\gamma(z_t) = \frac{p(x_1, ..., x_t, z_t)p(x_{t+1}, ..., x_T \mid z_t)}{p(X)} \tag{25}$$

$$\gamma(z_t) = \frac{\alpha(z_t)\beta(z_t)}{p(X)} \tag{26}$$

$$\tag{27}$$

donde

$$\alpha(z_t) \equiv p(x_1, ..., x_t, z_t) \tag{28}$$

$$\beta(z_t) \equiv p(x_{t+1}, ..., x_T \mid z_t) \tag{29}$$

$$\alpha(z_t) = p(x_t \mid z_t) \sum_{z_{t-1}} \alpha(z_t \mid z_{t-1}) \tag{30}$$

$$\alpha(z_1) = p(z_1)p(x_1 \mid z_1) = \prod_{k=1}^{K} \{\pi_k p(x_1 \mid \phi_k)\}^{z_{1k}} \tag{31}$$

Algoritmo backward-forward

$$\beta(z_t) = \sum_{z_{t+1}} \beta(z_{t+1}) p(x_{t+1} \mid z_{t+1}) p(z_{t+1} \mid z_t) \tag{32}$$

Contenido

Bootstrap

Contenido

Numero fijo de speakers

Numero variable de speakers

Resumen

- The first main message of your talk in one or two lines.
- The second main message of your talk in one or two lines.
- Perhaps a third message, but not more than that.

- Outlook
  - Something you haven't solved.
  - Something else you haven't solved.

Contenido

Trabajo futuro

Referencias I

A. Author.
Handbook of Everything.
Some Press, 1990.

S. Someone.
On this and that.
Journal of This and That, 2(1):50–100, 2000.