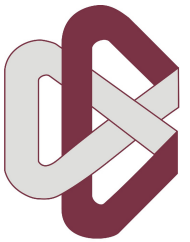
Segmentación de interlocutores a partir de señales de audio utilizando cadenas escondidas de Markov y técnicas de selección automática de modelos

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

1 Introducción

2 Speaker Diarization

3 Modelos de Markov

4 Selección de modelo

5 Metodología propuesta

6 Pruebas y resultados

7 Conclusiones

Contenido

1 Introducción
- Motivación
- Aplicaciones prácticas
- Contribuciones
- Principales enfoques
- Trabajo relacionado

2 Speaker Diarization

3 Modelos de Markov

4 Selección de modelo

5 Metodología propuesta

6 Pruebas y resultados

7 Conclusiones

Motivación

- En este trabajo de tesis abordarmos el siguiente problema:

  Se considera que una señal de audio contiene información de interés, y se desea encontrar el número de interlocutores y los segmentos de audio en los que participan.

- A este problema se le conoce en inglés como *Speaker Diarization*.

- Específicamente, se trata de:
  1. Encontrar el número total de personas que hablan en la conversación.
  2. Identificar los momentos en los que habla cada participante.

Aplicaciones prácticas

- 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.

Contribuciones

Las principales contribuciones de este trabajo se enumeran a continuación:

1. Se implementó el algoritmo Baum–Welch en Matlab para estimación de parámetros del HMM, con funciones críticas desarrolladas en C.

2. Se propuso un método de selección de modelo en dos etapas: primero usando una versión de BIC regularizada, y luego realizando pruebas de hipótesis a réplica bootstrap de un estadístico tipo LLR.

3. Se diseñó un sistema de pruebas para la simulación, parametrización y ajuste automático de múltiples modelos y luego la selección de mejor candidato.

4. Se realizaron pruebas con grabaciones de audio sintéticas que muestran el despempeño del método propuesto.

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 relacionado

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

1 Introducción

2 Speaker Diarization
- Formulación matemática
- Componentes del sistema
- Procesamiento acústico

3 Modelos de Markov

4 Selección de modelo

5 Metodología propuesta

6 Pruebas y resultados

7 Conclusiones

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

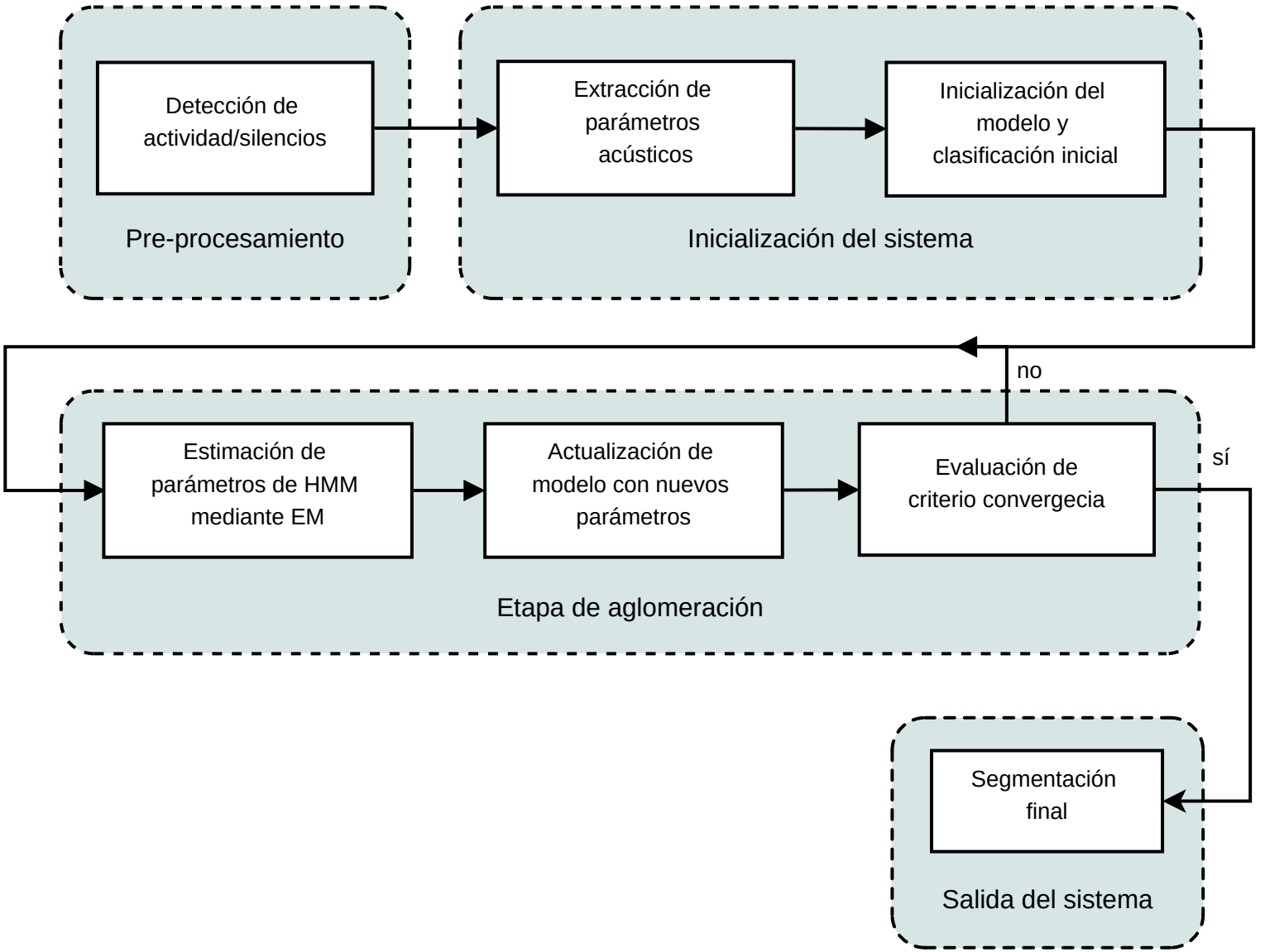
$$P(\mathcal{S}) = \Pi_{i=1}^{K} P(s_i \mid \phi(s_1, ..., s_{i-1})) \tag{6}$$

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 $\mathcal{A}$.

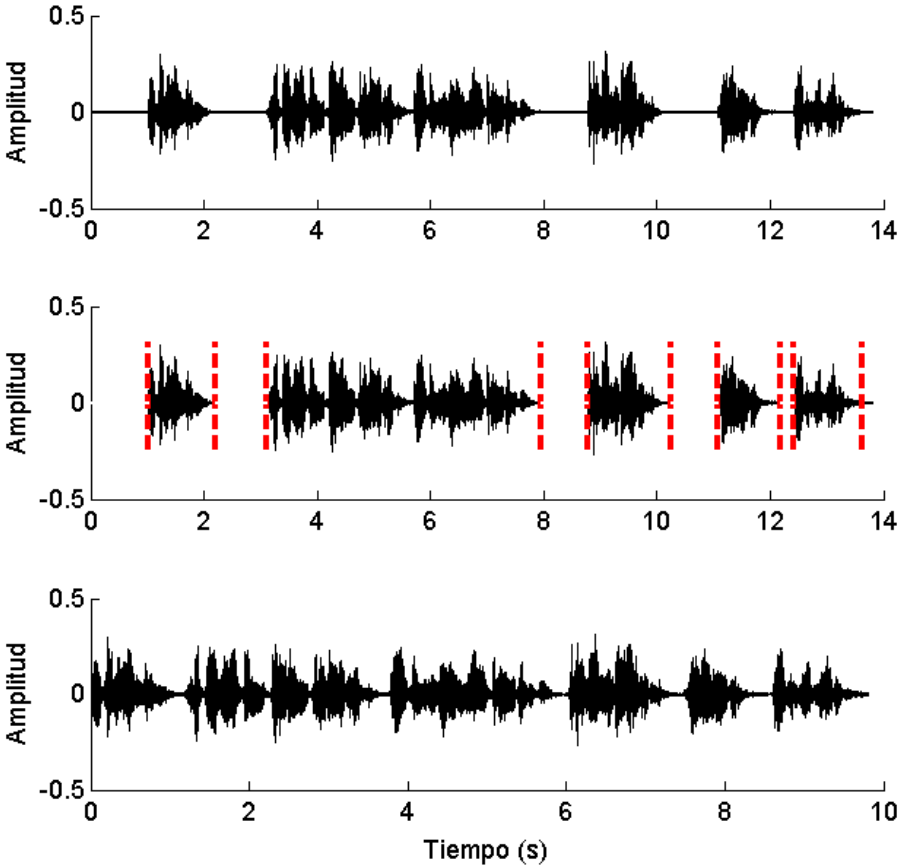
   Por otra parte, y después de obtener la segmentación correspondiente para la señal de audio, se deberá ahora inferir cuál es el modelo más probable de entre todos los que se propongan.

Componentes del sistema



Procesamiento acústico

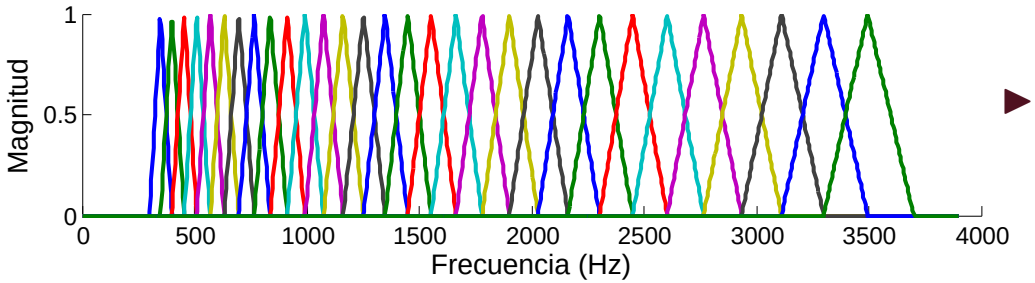
Pre-procesamiento de la señal



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

Procesamiento acústico

Obtención de vectores característicos (I)

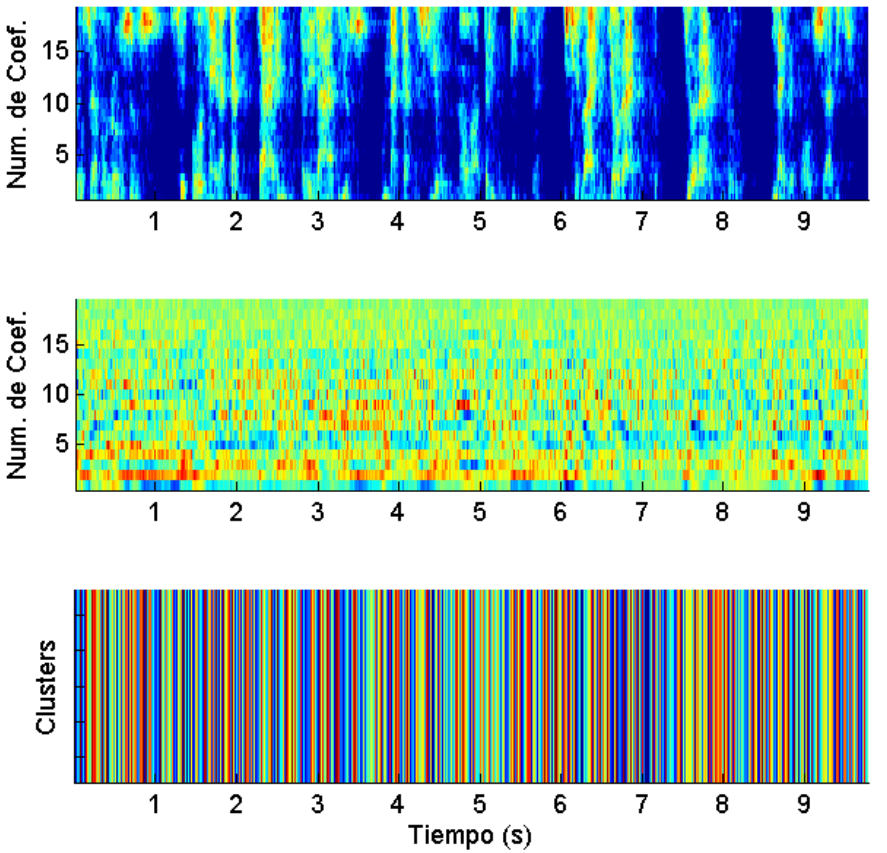


- Banco de filtros

- 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

1 Introducción

2 Speaker Diarization

3 Modelos de Markov
- Hidden Markov Model
- Resolver HMM con EM
- Algoritmo backward-forward

4 Selección de modelo

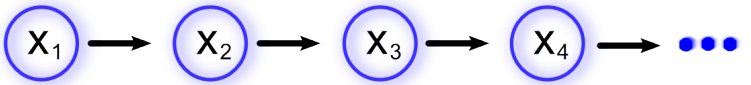
5 Metodología propuesta

6 Pruebas y resultados

7 Conclusiones

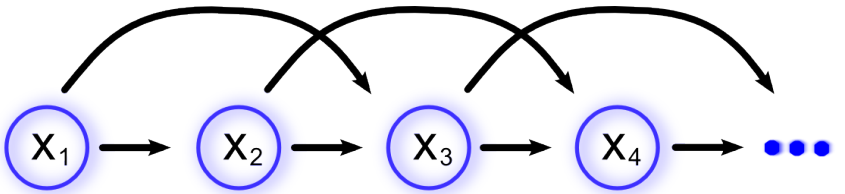
Cadenas de Markov

- 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



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

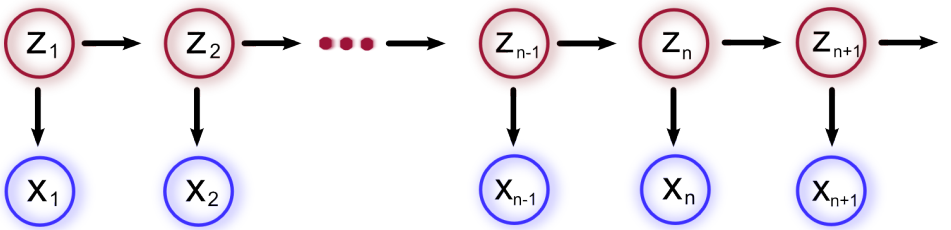
Modelo oculto de Markov

- 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 $\boldsymbol{\pi}$ 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.

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

Parámetros del HMM

- 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 son $\theta = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}, \phi\}$

- Es difícil determinar los parámetros del HMM por Máxima Verosimilitud, pues la función de verosimilitud es

$$p(\mathbf{X}, \theta) = \sum_{Z} p(\mathbf{X}, \mathbf{Z} \mid \theta) \qquad (17)$$

Estimación de parámetros de HMM con EM

- Si en vez de usar la función de verosimilitud (17), consideramos la verosimilitud de los datos completos $p(\mathbf{X}, \mathbf{Z})$, la maximización de esta función sería mucho más directa.
- No se conocen los datos completos $\{\mathbf{X}, \mathbf{Z}\}$.
- Utilizar el algoritmo EM para maximizar de forma eficiente la función de verosimilitud del HMM.
- Se propone usar la función de verosimilitud completa, y maximizar la esperanza del logaritmo de ésta.

$$\mathcal{Q}(\theta, \theta^{old}) = \mathbb{E}_{\mathbf{Z}|\theta^{old}} [\log p(\mathbf{X}, \mathbf{Z} \,|\, \theta)] \tag{18}$$

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{old}) \log p(\mathbf{X}, \mathbf{Z} \mid \theta) \tag{19}$$

Estimación de parámetros de HMM con EM

Notación

Se utilzará la siguiente notación para simiplificar el desarrollo:

Probabilidad marginal de una variable latente

$$\gamma(z_t) = p(z_t \mid \mathbf{X}, \theta^{old}) \tag{20}$$

Probabilidad conjunta de dos variables latentes consecutivas

$$\xi(z_{t-1}, z_T) = p(z_{t-1}, z_T \mid \mathbf{X}, \theta^{old}) \tag{21}$$

donde

$$\gamma(z_{tk}) = \mathbb{E}\left[z_{tk}\right] = \sum_{Z} \gamma(\mathbf{z}) \cdot z_{tk} \tag{22}$$

$$\xi(z_{t-1,j}, z_{tk}) = \mathbb{E}\left[z_{t-1,j} \cdot z_{tk}\right] = \sum_{Z} \gamma(\mathbf{z}) z_{t-1,j} \cdot z_{tk} \tag{23}$$

Estimación de parámetros de HMM con EM

E-Step

- Función de verosimilitud completa (reescrita con (22), (23))

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{t-1,j}, z_{tk}) \log A_{jk} + \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma(z_{tk}) \log p(x_T \mid \phi_k) \tag{24}$$

Estimación de parámetros de HMM con EM

M-Step

- $\theta_{new} = \underset{\theta}{arg\ max}\ \mathcal{Q}(\theta, \theta^{old})$
- Parámetros estimados por EM:

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})}, \quad A_{jk} = \sum_{t=2}^{T} \frac{\xi(z_{t-1,j}, z_{tk})}{\sum_{l=1}^{K} \xi(z_{t-1,j}, z_{tl})} \tag{25}$$

Algoritmo backward-forward

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

$$\gamma(z_t) = \frac{p(x_1, ..., x_t, z_t)p(x_{t+1}, ..., x_T \mid z_t)}{p(X)} \tag{27}$$

$$\gamma(z_t) = \frac{\alpha(z_t)\beta(z_t)}{p(X)} \tag{28}$$

$$\tag{29}$$

donde

$$\alpha(z_t) \equiv p(x_1, ..., x_t, z_t) \tag{30}$$

$$\beta(z_t) \equiv p(x_{t+1}, ..., x_T \mid z_t) \tag{31}$$

Algoritmo backward-forward

$$\alpha(z_t) = p(x_t \mid z_t) \sum_{z_{t-1}} \alpha(z_t \mid z_{t-1}) \tag{32}$$

$$\alpha(z_1) = p(z_1)p(x_1 \mid z_1) = \prod_{k=1}^{K} \{\pi_k p(x_1 \mid \phi_k)\}^{z_{1k}} \tag{33}$$

$$\beta(z_t) = \sum_{z_{t+1}} \beta(z_{t+1})p(x_{t+1} \mid z_{t+1})p(z_{t+1} \mid z_t) \tag{34}$$

$$\beta(z_T) = 1 \tag{35}$$

Factor de escala

- En el proceso recursivo para calcular cada $\alpha(z_n)$ se utilizan las probabilidades $\alpha(z_{n-1})$ anteriores...
- Manejar las probabilidades de forma re-escalada.

$$\hat{\alpha}(z_n) = p(z_n \,|\, x_1, ..., x_n) = \frac{\alpha(z_n)}{p(x_1, ..., x_n)} \tag{36}$$

- Se puede introducir un término de escala para cada $\hat{\alpha}(z_n)$

$$c_n = p(x_n \,|\, x_1, ..., x_{n-1}) \tag{37}$$

por lo que entonces

$$p(x_1, ..., x_n) = \prod_{i}^{n} c_i \tag{38}$$

Factor de escala

α

- A partir de (36) y (38) se puede escribir $\alpha(z_n)$ como sigue:

$$\begin{aligned}\alpha(z_n) &= p(x_1, ..., x_n, z_n) \\ &= \left(\prod_{i}^{n} c_i\right) \hat{\alpha}(z_n) \end{aligned} \tag{39}$$

y entonces la forma recursiva (32) se puede reescribir

$$c_n \hat{\alpha}(z_n) = p(x_n \,|\, z_n) \sum_{z_{n-1}} \hat{\alpha}(z_{n-1}) p(z_n \,|\, z_{n-1}) \tag{40}$$

Factor de escala

β

- Para re-escalar $\beta(z_n)$ usando los coeficientes c_n se tiene

$$\beta(z_n) = \left(\prod_{i}^{n} c_i \right) \hat{\beta}(z_n) \tag{41}$$

donde

$$\hat{\beta}(z_n) = \frac{p(x_{n+1}, ..., x_N \,|\, z_n)}{p(x_{n+1}, ..., x_N \,|\, x_1, ..., x_n)} \tag{42}$$

a partir de lo cual se puere escribir

$$c_{n+1} \hat{\beta}(z_n) = \sum_{z_{n+1}} \hat{\beta}(z_{n+1}) p(x_{n+1} \,|\, z_{n+1}) p(z_{n+1} \,|\, z_n) \tag{43}$$

Factor de escala

γ, ξ

- Y entonces tanto (20) como (21) se pueden formular de la siguiente manera:

$$\gamma(z_n) = \hat{\alpha}(z_n)\hat{\beta}(z_n) \tag{44}$$

$$\xi(z_{n-1}, z_n) = c_n \hat{\alpha}(z_{n-1}) p(x_n \mid z_n) p(z_n \mid z_{n-1}) \hat{\beta}(z_n) \tag{45}$$

Secuencia recuperada

- Como menciona Rabiner [**?**], hay muchas formas en las que se puede definir la secuencia óptima.
- Escoger las probabilidades marginales de las variables latentes $\gamma(z_n)$ (**??**) más probables de forma individual:

$$q_k = \underset{1<=k<=K}{arg\ max}\ \gamma(z_{nk}), \qquad 1 <= n <= N \tag{46}$$

- Otra opción, puede ser usar el algoritmo de Viterbi que nos permite obtener la secuencia de estados más probable dada una secuencia observada.

Contenido

1 Introducción

2 Speaker Diarization

3 Modelos de Markov

4 Selección de modelo
- BIC
- Bootsrap

5 Metodología propuesta

6 Pruebas y resultados

7 Conclusiones

Selección de modelo

Consideraciones

Hay varios aspectos importantes que considerar antes de abordar el problema de selección de modelos (Claeskens y Hjort [**?**]):

- Aproximación: La realidad observada suele ser mucho más compleja que los modelos propuestos.
- Sesgo-Varianza: Pocos parámetros a estimar, implican una menor variabilidad; mientras que más con modelos más complejos se reduce el sesgo.
- Parsimonia: 'El principio de parsimonia' o navaja de Ockham.
- Contexto: En algunos contextos puede ser más interesante encontrar los parámetros subyacentes del modelo e interpretarlos, mientras que en otros puede bastar con obtener respuesta a las problema planteado.

Funciones de penalización

- Una estrategia sencilla para la selección de modelo es elegir el candidato con la más grande probabilidad dados los datos.
- No siempre es un criterio lo suficientemente bueno para la comparación de modelos (sobre-ajuste del modelo).
- Utilizar una función que además de la verosimilitud del modelo, considere su complejidad

Criterio de Información Bayesiano

- Cuando hay varios modelos candidatos, una estategia bayesiana se encargaría de seleccionar el modelo que a posteriori sea más probable.
- Calcular la probabilidad posterior de cada uno de los modelos y luego seleccionando aquél modelo cuya probabilidad sea la mayor.
- Sean $\mathcal{M}_1, ..., \mathcal{M}_k$ los modelos propuestos, y sea $\mathbf{X} = \{x_1, ..., x_n\}$ el vector de datos observados. La probabilidad a posteriori para cada modelo se puede calcular como sigue:

$$P(\mathcal{M}_j \,|\, \mathbf{X}) \equiv \frac{P(\mathcal{M}_j)}{f(\mathbf{X})} \int_{\Theta_j} f(\mathbf{X} \,|\, \mathcal{M}_j, \theta_j) \pi(\theta_j \,|\, \mathcal{M}_j) d\theta_j \tag{47}$$

Criterio de Información Bayesiano

- La verosimilitud incondicional de los datos se puede calcular como sigue:

$$f(\mathbf{X}) = \sum_{j=1}^{k} P(\mathcal{M}_j)\lambda_{n,j}(y) \tag{48}$$

donde

$$\lambda_{n,j} = \int_{\Theta_j} \mathcal{L}_{n,j}(\theta_j)\pi(\theta_j \mid \mathcal{M}_j)d\theta_j. \tag{49}$$

- La ecuación (49) representa la verosimilitud marginal de los datos para el modelo $\mathcal{M}_j$ integrada con respecto a θ_j sobre es espacio de parámetros Θ_j correspondiente.

Criterio de Información Bayesiano

- Ahora, si se define

$$BIC_{n,j}^{exact} \equiv 2log(\lambda_{n,j}(\mathbf{X})) \tag{50}$$

por lo que (47) se podría reescribir como sigue:

$$P(\mathcal{M}_j \mid \mathbf{X}) = \frac{P(\mathcal{M}_j)exp(\frac{1}{2}BIC_{n,j}^{exact})}{\sum_{i=1}^{k} P(\mathcal{M}_i)exp(\frac{1}{2}BIC_{n,i}^{exact})} \tag{51}$$

- El cálculo de los diferentes $BIC_{n,j}^{exact}$ es difícil de estimar numéricamente, además de que se necesitan las probabilidades a priori para todos los modelos y todos los parámetros; por lo que se usará una expresión similar que sea práctica y mucho más eficiente.

Bootstrap

- Bootstrap es una técnica estadística que nos permite tener noción sobre qué tan precisa es alguna medida muestral estimada.
- Permite aproximar la distribución de muestreo de casi cualquier estadístico, usando métodos sencillos pero computacionalmente intensivos.
- Como menciona Persi et.al [**?**], esta técnica fue desarrollada en 1978 por Efron [**?**], quien generalizó el método de *Jacknife*.

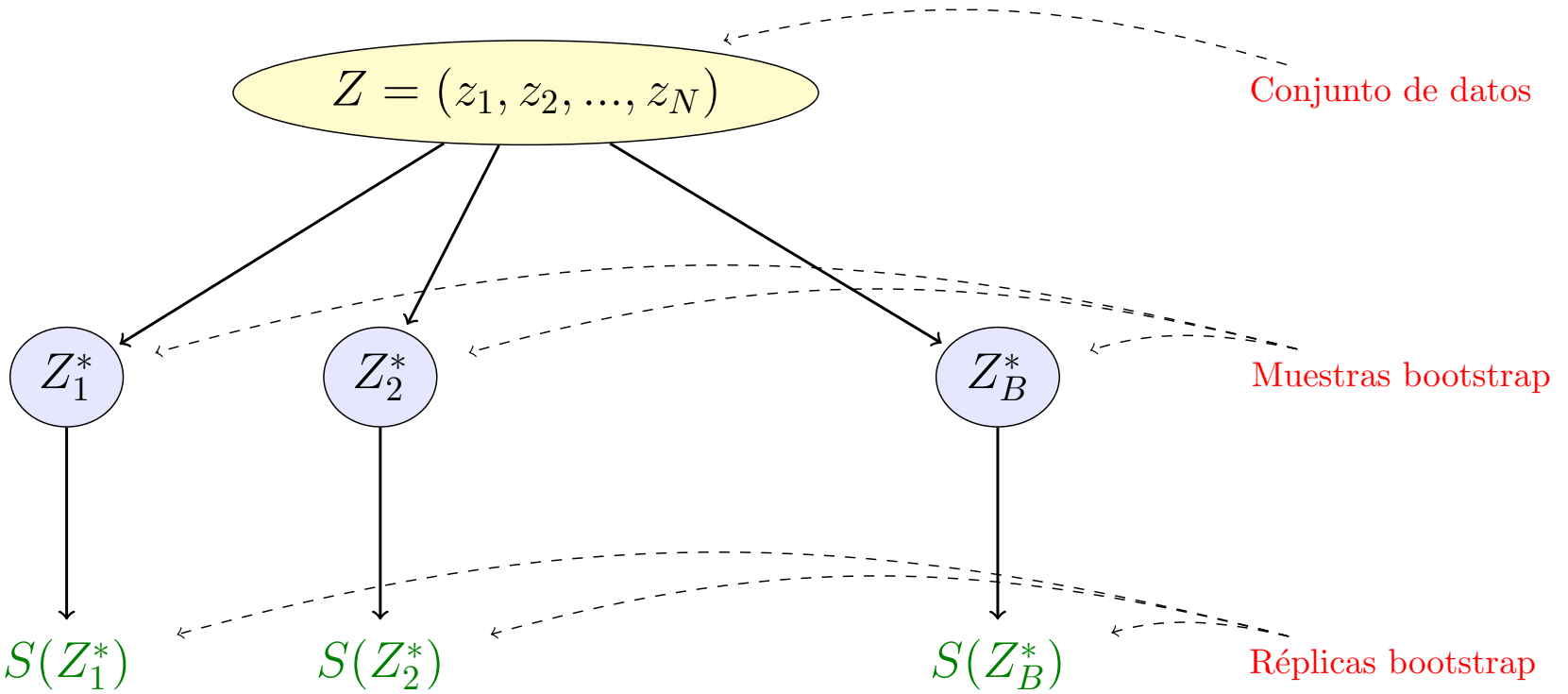
Bootstrap no paramétrico

- Se tiene que ajustar un modelo a un conjunto de datos. Sea este conjunto de entrenamiento $\mathbf{Z} = (z_1, z_2, ..., z_N)$ donde $z_i = (x_i, y_i)$ y son independientes con distribución F.
- Como $\mathbf{Z}$ es una muestra finita no se conoce tal cual la distribución F,
- Se estimarpa una función empírica $\hat{F}$ donde a cada observación z_i se le asigna un peso $\frac{1}{N}$ en la densidad.

Bootstrap no paramétrico

- Seleccionar de forma aleatoria y con reemplazo de $\hat{F}$ un conjunto de datos del mismo tamaño que el conjunto original.
- A este conjunto se le denotará $\mathbf{Z}_1^*$. Este proceso de selección se realiza B veces, produciendo B conjuntos bootstrap $\mathbf{Z}_\cdot^* = \{\mathbf{Z}_1^*, \mathbf{Z}_2^*, ..., \mathbf{Z}_B^*\}$.
- Para cada uno de estos conjuntos resultantes, se volvera a ajustar el modelo, y se examinará el comportamiento de los ajustes para las respuestas obtenidas, obteniendo lo que se conoce como réplica bootstrap.

Bootstrap no paramétrico



Bootsrap paramétrico

- El bootstrap clásico es no paramétrico, y se vale únicamente del conjunto observado para a partir de ahí estimar la función de distribución empírica.
- En el caso que nos interesa, se cuenta con un modelo paramétrico que ha sido ajustado a los datos, usualmente por MLE.
- A partir de este modelo ajustado es que se muestrea. Al igual que en con la técnica no paramétrica, se suelen generan muestras de datos del mismo tamaño que el conjunto original.
- Luego, para cada nueva conjunto bootstrap $\mathbf{Z}_b^*$ muestreado se calcula el estadístico de nuestro interés. Éste proceso de muestreo se repite igualmente una gran cantidad de veces.

Selección de modelo usando BIC

- Puesto que consideramos que dentro de nuestro espacio de modelos propuestos se encuentra el modelo solución, resulta más natural usar BIC (cita).
- Esta función de penalización se calcula de la siguiente manera:

$$BIC(\mathcal{M}) = 2\mathcal{L}_{max}(\mathcal{M}) - (logN)dim(M) \tag{52}$$

  para cada modelo $\mathcal{M}$, donde $dim(M)$ es el número estimado de parámetros libres que le corresponden, y N es el tamaño de nuestra muestra de datos.
- Por otra parte, $\mathcal{L}_{max}(\mathcal{M})$ es la máxima log-verosimilitud obtenida para el modelo $\mathcal{M}$ después de realizar un número HMM_{MAX} de simulaciones, para evitar que en algún caso el algoritmo de estimación se quede atorado en un máximo local.

Selección de modelo usando BIC

- Para estimar el número de parámetros libres de nuestro modelo, se consideran todas las probabilidades que rigen al HMM:
  - Matriz a priori o inicial
  - Matriz de transición entre los interlocutores
  - Matriz de emisión de cada persona para todo el diccionario de palabras
- Este tipo de función de penalización nos permite seleccionar de entre un conjunto de modelos propuestos (que pueden ser muchos) al modelo o los modelos con mayor probabilidad de ser los correctos.

Selección de modelo usando bootstrap con likelihood ratio testing

- Se utilizará la técnica bootstrap paramétrico, pues mediante el algoritmo EM es fácil obtener el modelo parametrizado.
- Usualmente se compararán dos modelos adyacentes, es decir, el modelo $\mathcal{M}_d$ de d estados contra el modelo $\mathcal{M}_{d+1}$ de $d+1$ estados ocultos.
- Se usa el estadístico LLR que corresponde a la diferencia de las log-verosimilitudes de dos modelos

$$LLR_{obs}^{(d)} = \log \frac{L(\hat{\theta}^{(d+1)}; y_{1:n})}{L(\hat{\theta}^{(d)}; y_{1:n})} = \log L(\hat{\theta}^{(d+1)}; y_{1:n}) - \log L(\hat{\theta}^{(d)}; y_{1:n}) \quad (53)$$

Selección de modelo usando bootstrap con likelihood ratio testing

- Para calcular el MLE de un modelo, se estimó la verosimilitud varias veces con diferentes parámetros iniciales aleatorios.
- Se iteró el algoritmo EM hasta convergencia, un numero $iter_{hmm}$ fijo de iteraciones, esto para evitar el estancamiento del algoritmo en un máximo local, y obtener así una buena estimación de la máxima verosimilitud del modelo.
- Se usó entonces como MLE del modelo la máxima verosimilitud correspondiente a los mejores parámetros estimados y que se denotó por LLR_{obs}.

Contenido

1 Introducción

2 Speaker Diarization

3 Modelos de Markov

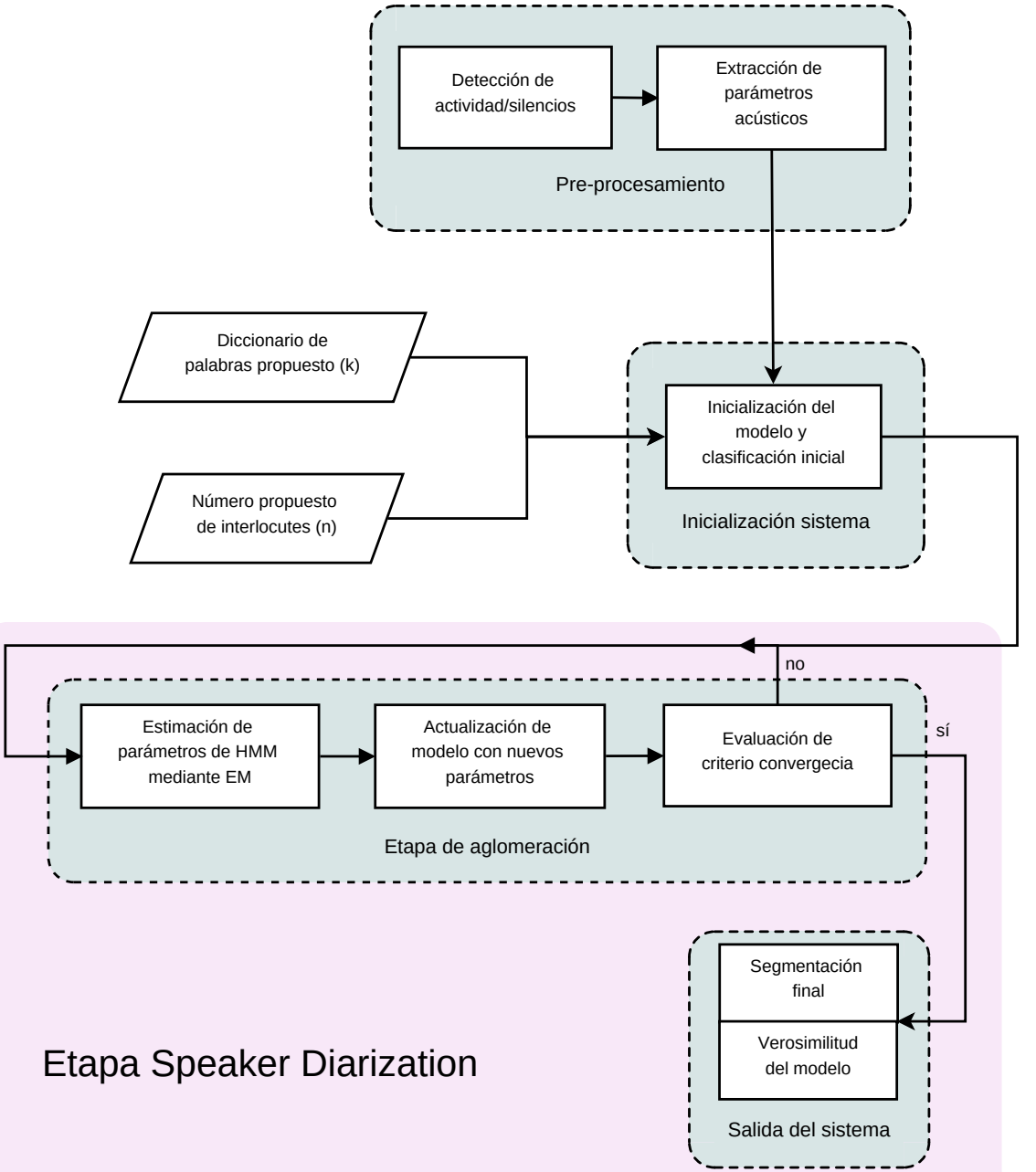
4 Selección de modelo

5 Metodología propuesta

6 Pruebas y resultados

7 Conclusiones

Metodología propuesta



Selección de modelo realizada

Exploración

- Se estimará BIC de todos los HMM propuestos para generar una curva de selección.
- Como se verá en las pruebas, es necesario introducir un término de regularización en BIC para que la penalización del modelo corresponda con las log-verosimilitudes obtenidas.

$$BIC_{\lambda}(\mathcal{M}) = 2\mathcal{L}_{max}(\mathcal{M}) - \lambda \cdot logN \cdot dim(M) \tag{54}$$

- Se deberá encontrar el valor adecuado para λ que penalice de buena forma los modelos.

Selección de modelo realizada

Exploración

- Para escoger el valor adecuado de λ se realizará un análisis de sensibilidad, generando múltiples curvas de selección BIC con diferentes valores de λ.
- Se buscará la región de inflexión que divide a la superficie anterior en dos: en la primera parte de la superficie.
- Calculando el gradiente de la superficie generada por las funciones BIC, se buscará la región asociada a un λ en el que la suma de los valores absolutos sea menor.
- Una vez que se logre seleccionar el λ adecuado de regularización, se procede a evaluar su curva BIC asociada y de ahí se obtiene al modelo ganador o un subconjunto de posibles ganadores.

Selección de modelo realizada

Selección

- Luego se realizará un proceso de refinamiento en caso de que se tengan varios modelos posibles.
- Se formarán pares de modelos que se deseen comparar, y se estimará su LLR, que se denominará como LLR_{obs}.
- Luego, mediante bootstrap paramétrico se hará una prueba de hipótesis para comproar cuál modelo es más adecuado para los datos.
- De esta forma, se pueden realizar pruebas de hipótesis para los modelos candidatos, e ir rechazando modelos de acuerdo al análisis propuesto.

Contenido

1 Introducción

2 Speaker Diarization

3 Modelos de Markov

4 Selección de modelo

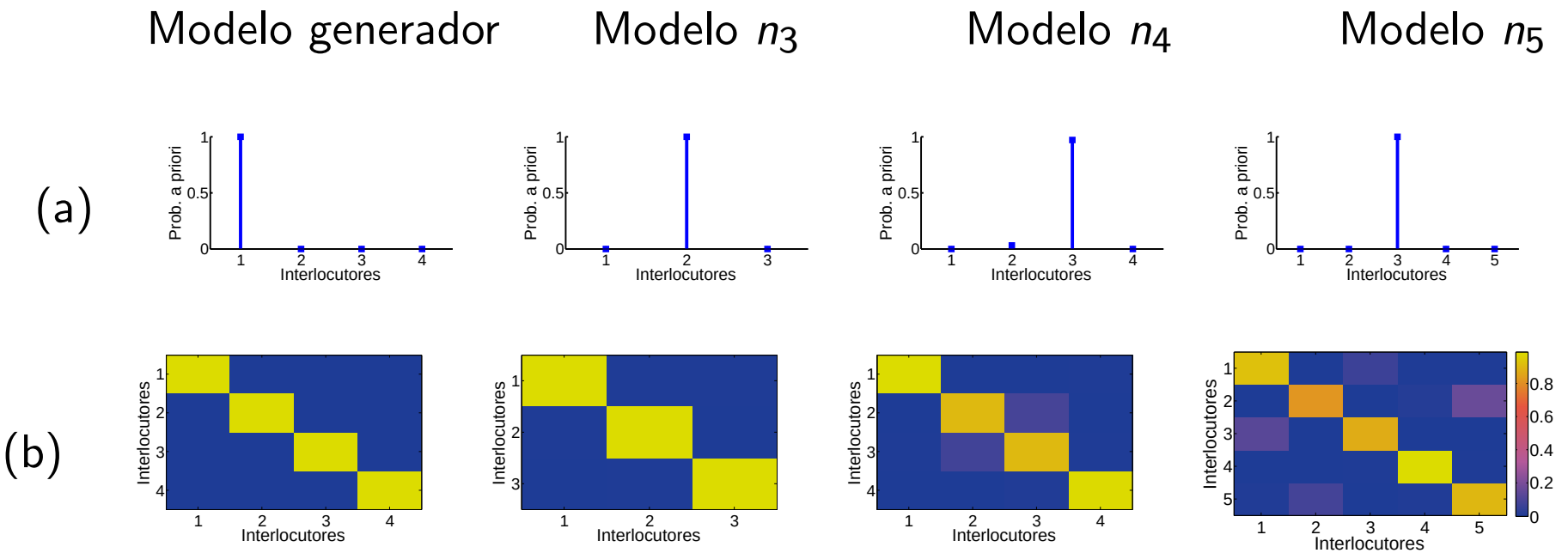
5 Metodología propuesta

6 Pruebas y resultados
- Pruebas
- Resultados

7 Conclusiones

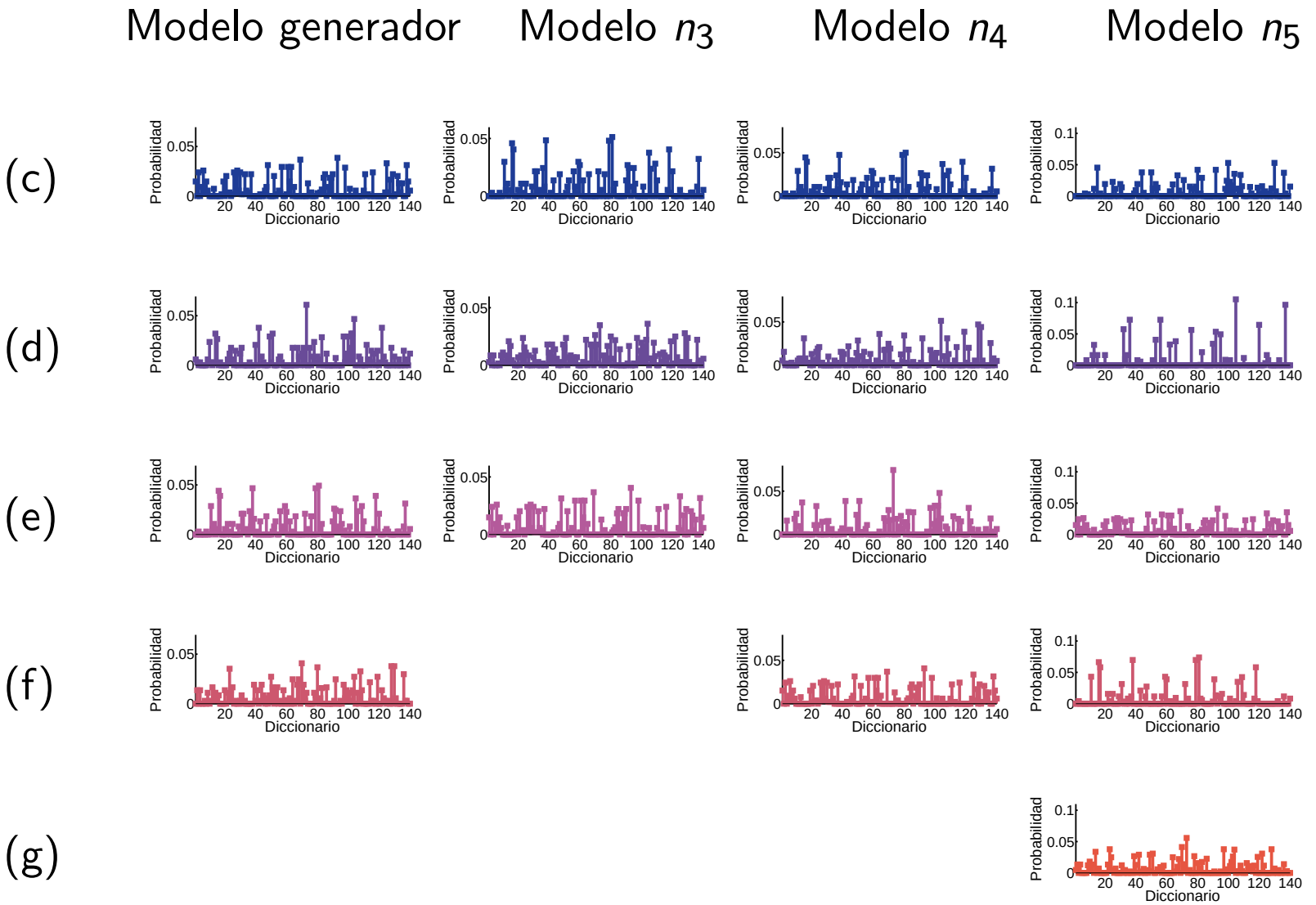
Secuencia 3

Parámetros



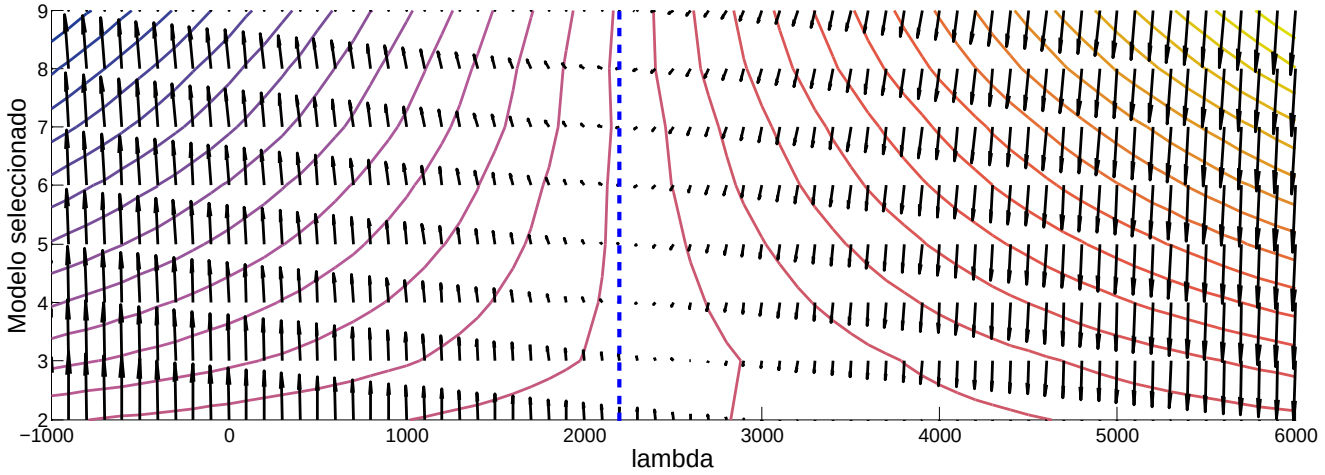
Secuencia 3

Parámetros

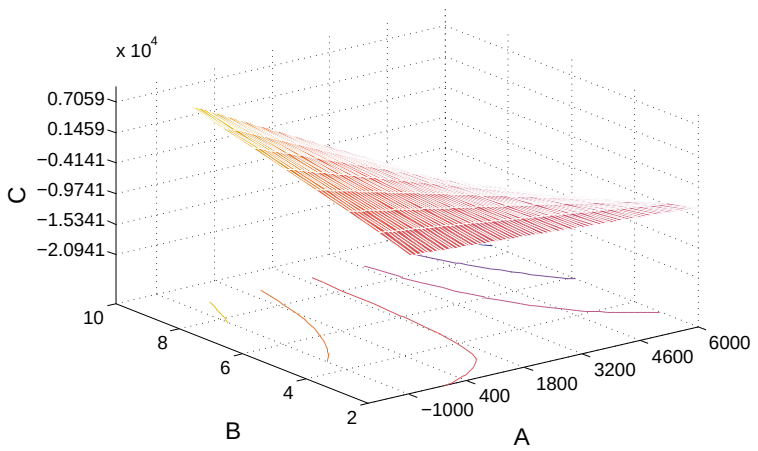


Secuencia 3

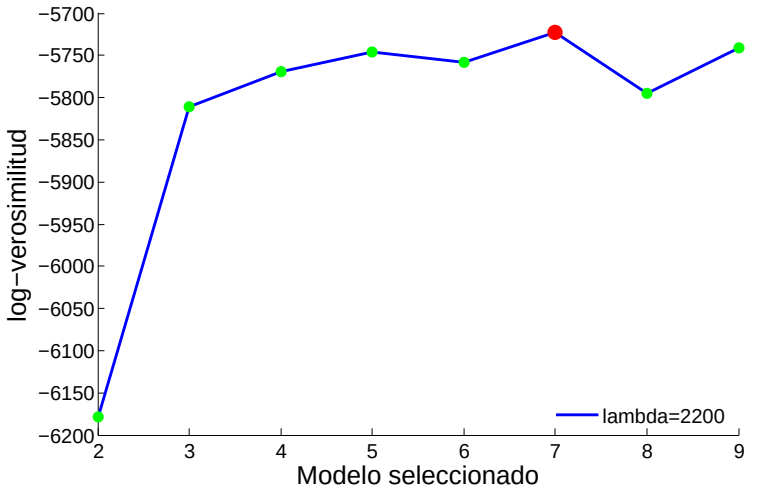
Superficie BIC



(a)



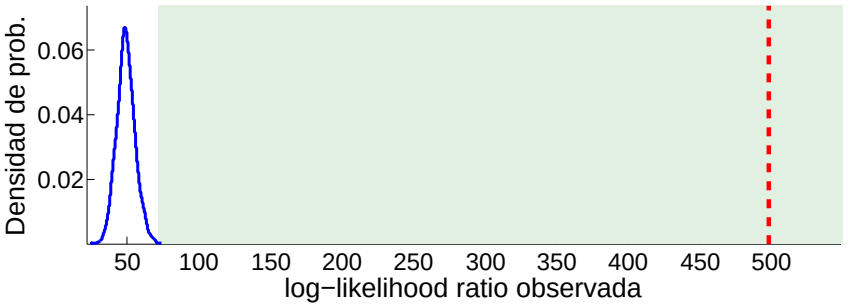
(b)



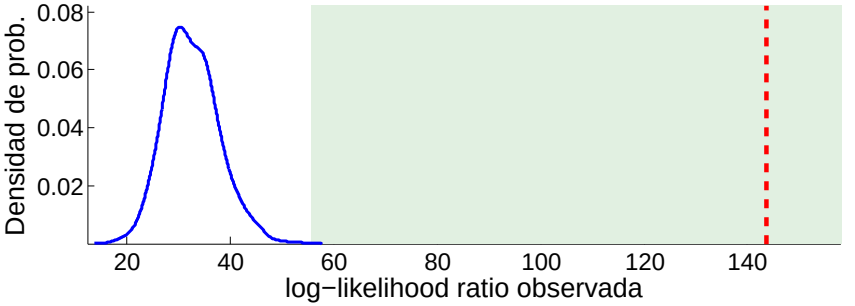
(c)

Secuencia 3

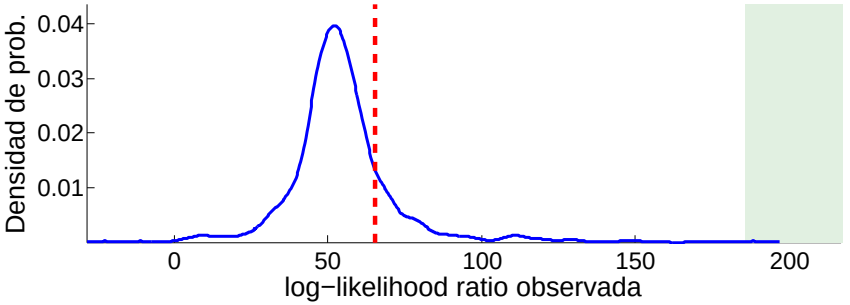
Pruebas de hipótesis



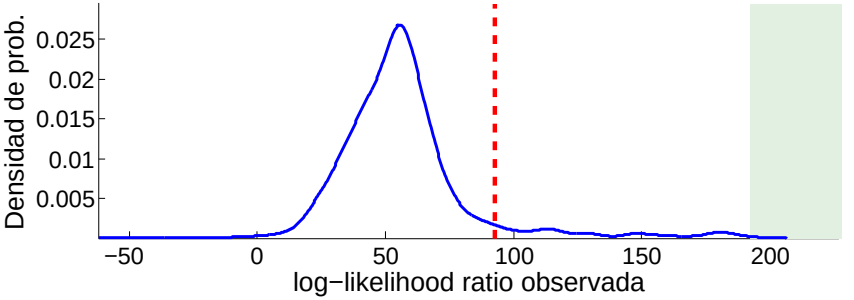
(a)



(b)



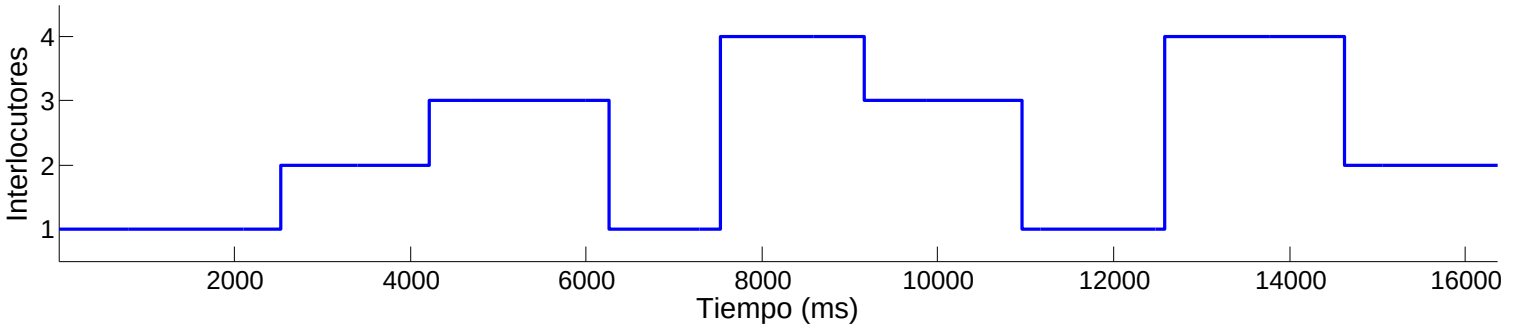
(c)



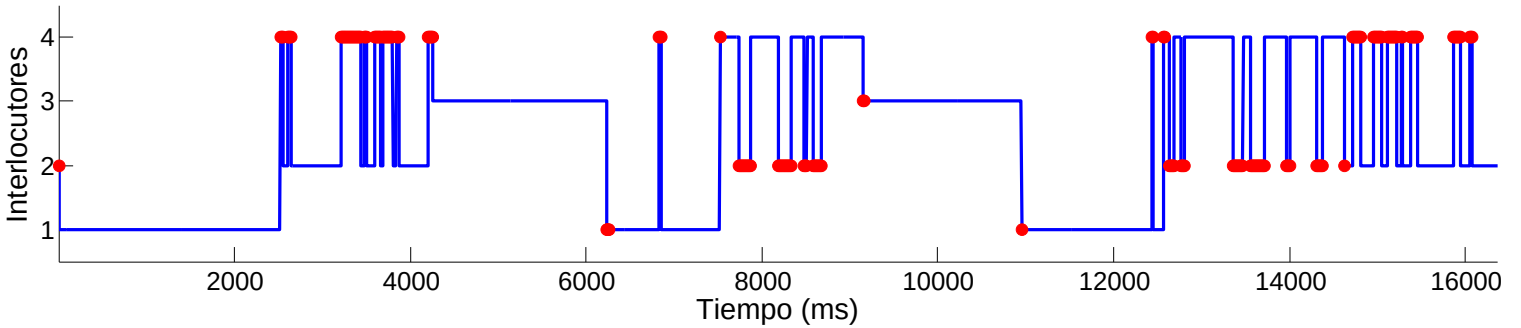
(d)

Secuencia 3

Secuencia recuperada



(a)



(b)

Tabla de resultados (I)

Descripción de las secuencias de audio utilizadas para las pruebas.

SEC. ID.	DURACIÓN	#SAMPLES	#WORDS	n_{TRUE}
ALLPOE	12:06 min	7218	140	6
GARMÁR	6:55 min	4154	90	4
WSHAKE	2:43 min	1636	160	4
MACUÑA	10:41 min	6414	120	3
CALDER	1:07 min	676	160	3
LLOYDW	6.48 min	4084	160	5

Tabla 1: *Descripción de características de secuencias utilizadas para las pruebas*

Tabla de resultados (II)

Detalle de los resultados para todas las secuencias de audio.

SEC. ID.	n_{TRUE}	n_{FOUND}	DER (%)
ALLPOE	6	6	03.324
GARMÁR	4	4	00.652
WSHAKE	4	4	12.353
MACUÑA	3	3	00.343
CALDER	4	3	*08.092
LLOYDW	5	5	01.170

Tabla 2: *Resultados de pruebas realizadas. Se observa en general que para la mayoría de las pruebas se encontró el modelo correcto, además de que se tiene un DER en general abajo del 10 %. En el caso de la secuencia* `CALDER`*, se muestra el DER correspondiente al modelo correcto.*

Contenido

1 Introducción

2 Speaker Diarization

3 Modelos de Markov

4 Selección de modelo

5 Metodología propuesta

6 Pruebas y resultados

7 Conclusiones

Conclusiones

1. De acuerdo a los resultados obtenidos, se observa que el método presentado muestra un buen desempeño en las pruebas realizadas.

2. La metodología propuesta permite una rápida exploración de todos los modelos propuestos, así como una selección del mejor candidato de acuerdo a las pruebas de hipótesis que se plantean.

3. La calidad de los resultados depende en gran parte de un correcto pre-procesamiento de la señal; ya sea para eliminar ruidos, así como para la correcta parametrización de los vectores característicos.

4. Aunque el método presentado realiza pruebas computacionalmente intensivas, desde la primera etapa de exploración permite identificar a un sub-conjunto pequeño de posibles modelos correctos.

5. La importancia de la segunda etapa de selección, permite dar certeza sobre cuál de los modelos es el correcto. Incluso con un alto nivel significancia, las pruebas de hipótesis suelen seleccionar de buena forma al modelo ganador.

Trabajo futuro

1. Construir un banco de pruebas con voces reales, que permitan analizar el comportamiento del método presentado en un entorno más real.
2. Mejorar la forma en que se eliminan los silencios y ruidos; pues en pruebas con un ambiente *normal*, hay muchas más fuentes de perturbación que las que hasta ahora se han considerado.
3. Explorar otras estrategias para selección de la segmentación más óptima, buscando realizar el cálculo de manera eficiente.
4. Evaluar la pertinencia de paralelizar el algoritmo principal de estimación de parámetros del HMM, para mejorar el rendimiento general del sistema.

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.