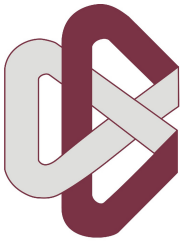
Segmentación de interlocutores a partir de señales de audio utilizando cadenas escondidas de Markov y técnicas de selección automática de modelos

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

- 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.

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 la segmentación correcta 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 (II)

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:

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

Que por el Teorema de Bayes, 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

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

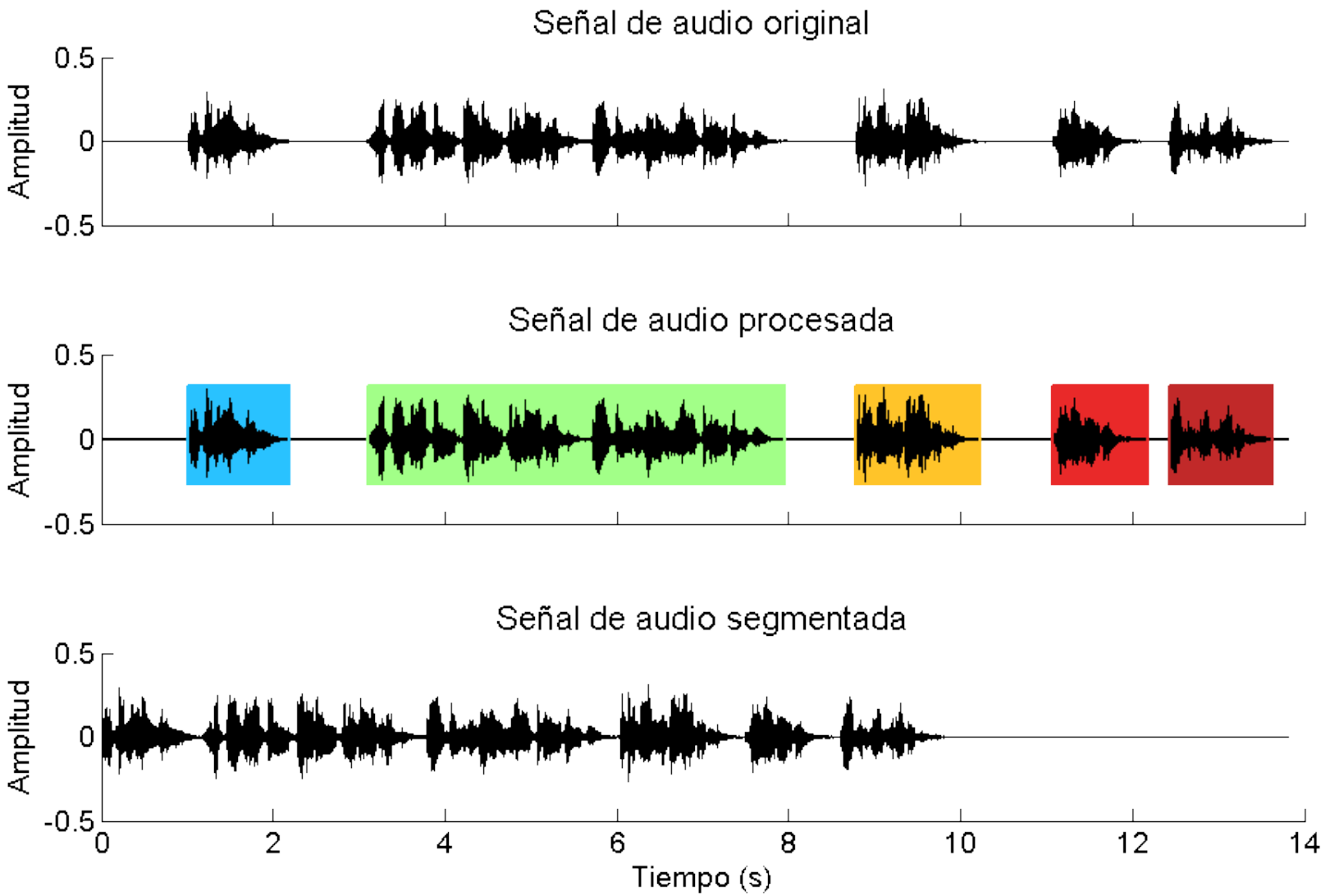
1. Procesamiento acústico: Se refiere a la forma en la que se procesará la información y se digitalizará.Además se debe realizar algún proceso para obtener una representación paramétrica de la señal. A este procedimiento se le conoce como construcción del *diccionario de palabras*.

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)

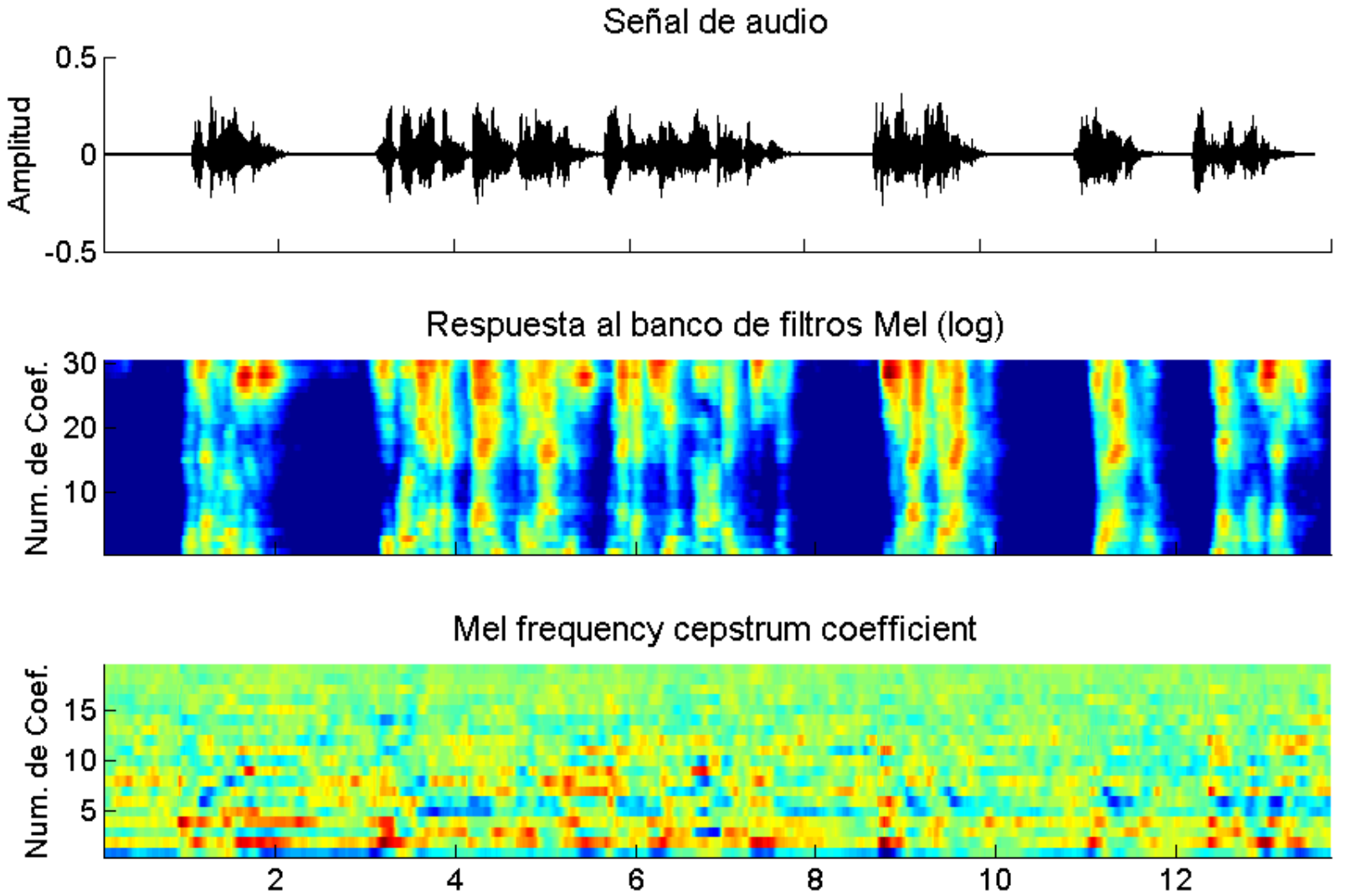
1. Modelado de interlocutores:
2. Búsqueda de hipótesis:

Elminación de ruido / Detección de silencios



Mel Frequency Cepstrum Coefficient

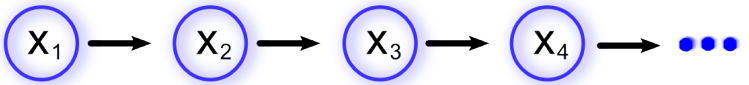
- FFT (ventana) -> Banco de filtros triangular (Mel Scale) -> Log -> DCT -> MFCC



Contenido

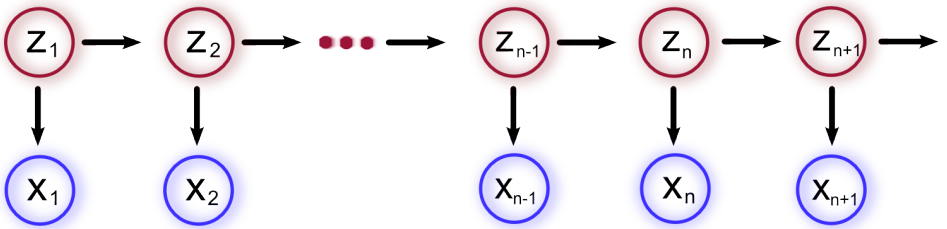
Hidden Markov Model

- Cadena de Markov de primer orden



$$p(x_1, ..., x_N) = \prod_{n=1}^{N} p(x_n \mid x_1, ..., x_{n-1}) = p(x_1) \prod_{n=2}^{N} p(x_n \mid x_{n-1}) \quad (5)$$

- Modelo estocástico de Markov en el que el estado de la cadena es parcialmente observado.



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

Hidden Markov Model

- Agregar una variable latente z_n (discreta), que corresponda a cada observación x_n.

$$z_{n+1} \perp z_{n-1} \mid z_n \tag{6}$$

$$p(x_1, ..., x_N, z_1, ..., z_N) = p(z_1) \left[\prod_{n=2}^{N} p(z_n \mid z_{n-1}) \right] \prod_{n=1}^{N} p(x_n \mid z_n). \tag{7}$$

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

Parámetros del HMM

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

$$A_{jk} \equiv p(z_{nk} = 1 \mid z_{n-1,j} = 1) \tag{8}$$

$$p(z_n \mid z_{n-1}, \mathbf{A}) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j} \cdot z_{n,k}} \tag{9}$$

- Vector de distribución inicial π para variable latente.

$$\pi_k \equiv p(z_{1k}) \tag{10}$$

$$p(z_1 \mid \pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \tag{11}$$

- Probabilidad de emisión de una variable observada x_n dada una variable latente z_n.

$$p(x_n \mid z_n, \phi) = \prod_{k=1}^{K} p(x_n \mid \phi_k)^{z_{nk}} \tag{12}$$

HMM con EM

- Probabilidad conjunta del modelo

$$p(\mathbf{X}, \mathbf{Z} \mid \theta) = p(z_1 \mid \pi) \left[\prod_{n=2}^{N} p(z_n \mid z_{n-1}, \mathbf{A}) \right] \prod_{n=1}^{N} p(x_n \mid z_n, \phi) \tag{13}$$

- Parámetros del modelo $\theta = \{\boldsymbol{\pi}, \mathbf{A}, \phi\}$
- Función de verosimilitud completa

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \theta^{old}) \log p(\mathbf{X}, \mathbf{Z} \mid \theta) \tag{14}$$

- Prob. marginal de una variable latente, prob. conjunta de dos variables latentes consecutivas

$$\gamma(z_n) = p(z_n \mid \mathbf{X}, \theta^{old}) \tag{15}$$

$$\xi(z_{n-1}, z_n) = p(z_{n-1}, z_n \mid \mathbf{X}, \theta^{old}) \tag{16}$$

HMM con EM

- Prob. marginal de $z_n k = 1$, prob. conjunta de $z_{n-1,j}, z_{nk}$

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \sum_{Z} \gamma(\mathbf{z}) z_{nk} \tag{17}$$

$$\xi(z_{n-1,j}, z_{nk}) = \mathbb{E}[z_{n-1,j} \cdot z_{nk}] = \sum_{Z} \gamma(\mathbf{z}) z_{n-1,j} \cdot z_{nk} \tag{18}$$

- Función de verosimilitud completa (reescrita con (17), (18))

$$\begin{aligned} \mathcal{Q}(\theta, \theta^{old}) = \sum_{k=1}^{K} \gamma(z_{1k}) \log \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \log A_{jk} + \\ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \log p(x_n \mid \phi_k) \end{aligned} \tag{19}$$

- Parámetros estimados por EM:

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_1 j)}, \quad A_{jk} = \sum_{n=2}^{N} \frac{\xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \xi(z_{n-1,j}, z_{nl})} \tag{20}$$

Algoritmo backward-forward

$$\gamma(z_n) = p(z_n \mid X) = \frac{p(X \mid z_n)p(z_n)}{p(X)} \tag{21}$$

$$\gamma(z_n) = \frac{p(x_1, ..., x_n, z_n)p(x_{n+1}, ..., x_N \mid z_n)}{p(X)} \tag{22}$$

$$\gamma(z_n) = \frac{\alpha(z_n)\beta(z_n)}{p(X)} \tag{23}$$

$$\tag{24}$$

donde

$$\alpha(z_n) \equiv p(x_1, ..., x_n, z_n) \tag{25}$$

$$\beta(z_n) \equiv p(x_{n+1}, ..., x_N \mid z_n) \tag{26}$$

$$\alpha(z_n) = p(x_n \mid z_n) \sum_{z_{n-1}} \alpha(z_n \mid z_{n-1}) \tag{27}$$

$$\alpha(z_1) = p(z_1)p(x_1 \mid z_1) = \prod_{k=1}^{K} \{\pi_k p(x_1 \mid \phi_k)\}^{z_{1k}} \tag{28}$$

Algoritmo backward-forward

$$\beta(z_n) = \sum_{z_{n+1}} \beta(z_{n+1}) p(x_{n+1} \mid z_{n+1}) p(z_{n+1} \mid z_n) \tag{29}$$

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.