# 11. IDENTIFICACIÓN DE LOCUTORES

22.47 Procesamiento de voz

Marc S. Ressl

# Identificación de locutores

El reconocimiento de voz es un ingrediente fundamental en la implementación de agentes virtuales.

Pero sólo nos dice **qué** se está diciendo; no nos dice **quién** lo está diciendo.

# Identificación de locutores



La **verificación de locutores** determina si una grabación de voz es de quien dice ser. Consiste en una clasificación 1:1.

En **identificación de locutores** asignamos una identidad a una grabación de voz. Consiste en una clasificación 1:N, o, si se admiten locutores desconocidos, 1:N+1.

Ambos problemas pertenecen al campo de la **biometría**.

# Identificación de locutores



Otras aplicaciones de la verificación/identificación de locutores:

- Sistemas de control de acceso.
- Sistemas forenses.
- Anotación automática de conversaciones (big data).

# Enfoque clásico

# Features

Al igual que en reconocimiento de voz, convertimos la señal de voz $s[n]$ es una secuencia de feature vectors $X = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$, en donde cada feature vector es $\mathbf{X}_t = \{x_{t1}, x_{t2}, \ldots, x_{tD}\}$.

Necesitamos features que sean fuertemente dependientes de la **estructura física del aparato fonador** y de los **hábitos articulatorios** de un locutor.

# Features

En la práctica solemos usar los **mfcc**, por más que descartan información útil en identificación de locutores.

También usamos los **Δmfcc** (la diferencia entre vectores mfcc sucesivos) y los **Δ²mfcc** (la diferencia entre vectores Δmfcc sucesivos).

# MAP

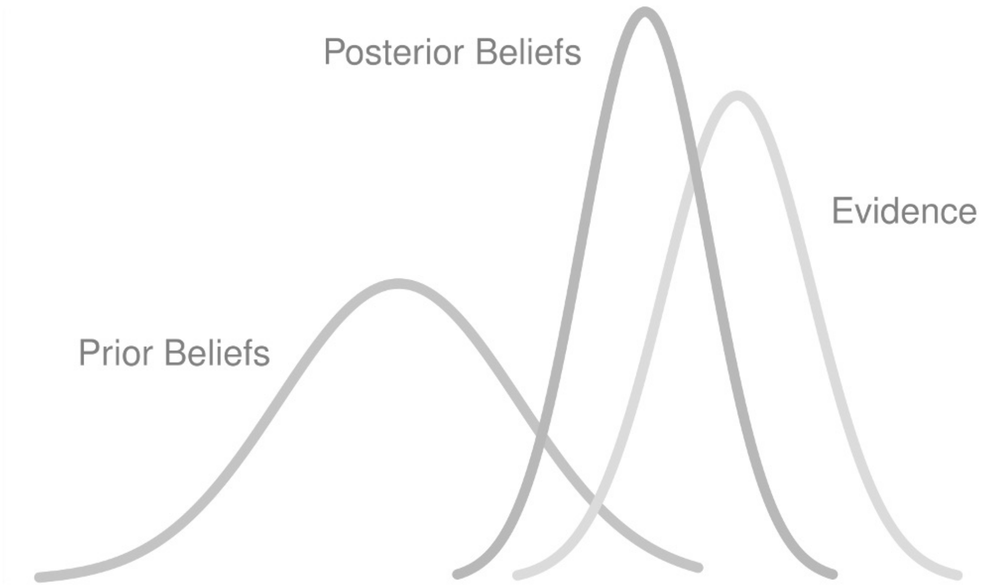
Para determinar la identidad de un locutor solemos aplicar el criterio *maximum a posteriori* (MAP):

$$\hat{L} = \arg\max_{L} P(L \mid X)$$

$\hat{L}$ es el locutor que maximiza la probabilidad MAP.

$L$ es el espacio de posibles locutores.

$X$ es la matriz de feature vectors.



# MAP

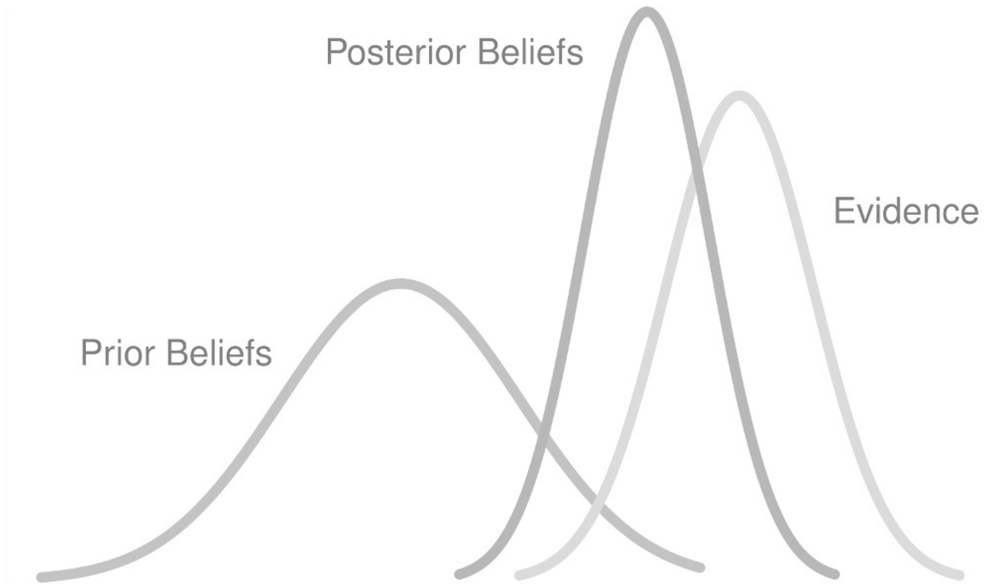
Por la regla de Bayes:

$$\hat{L} = \arg\max_{L} \frac{P(X \mid L)P(L)}{P(X)}$$

$P(X|L)$ es el **modelo del locutor**.

$P(L)$ es la **probabilidad de ocurrencia de locutor**.

$P(X)$ no depende de $L$, por lo que solemos ignorarlo.



# GMM

Para modelar $P(X|L)$ solemos usar modelos **GMM** (gaussian mixture models).

Solemos ignorar el aspecto temporal de la matriz de feature vectors $X = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$, considerando los valores $\mathbf{X}_t$ como muestras de un proceso i.i.d.

# GMM

Una **gaussiana multidimensional** se define como:

$$p_{\mu,\Sigma}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

$\mu$ es un vector de medias.

$\Sigma$ es una matriz de covarianza, $|\Sigma|$ su determinante.

$d$ es la dimensión de los feature vectors.

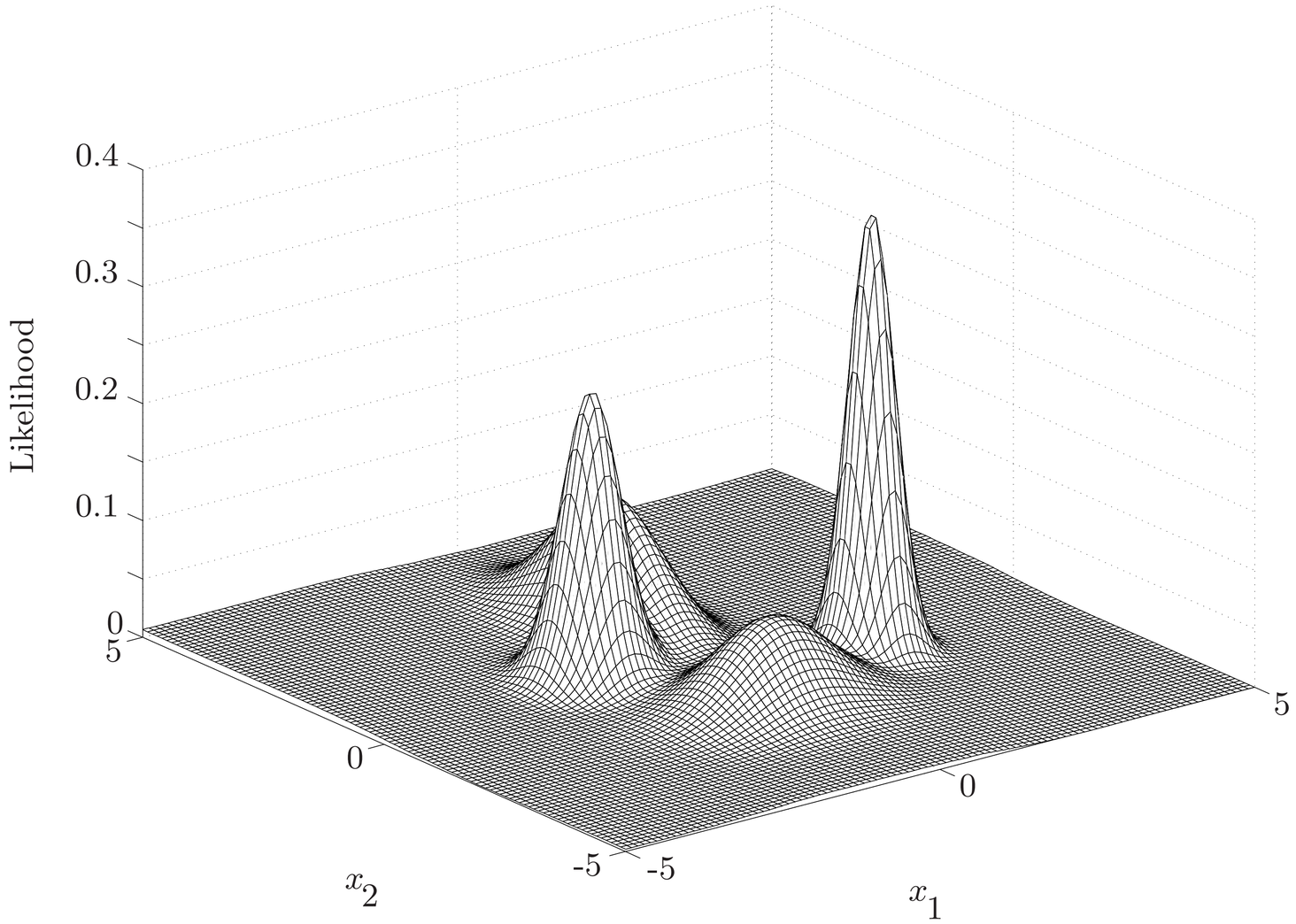
# GMM

El **modelo de mezclas gaussianas** se construye a partir de una combinación lineal de $N$ gaussianas:

$$P(X \mid L) = \sum_{k=1}^{N} w_k p_{\mu_k, \Sigma_k}(X)$$

Los pesos $w_k$ han de cumplir:

$$\sum_{k=1}^{N} w_k = 1$$




# Modelo GMM

$N = 4$

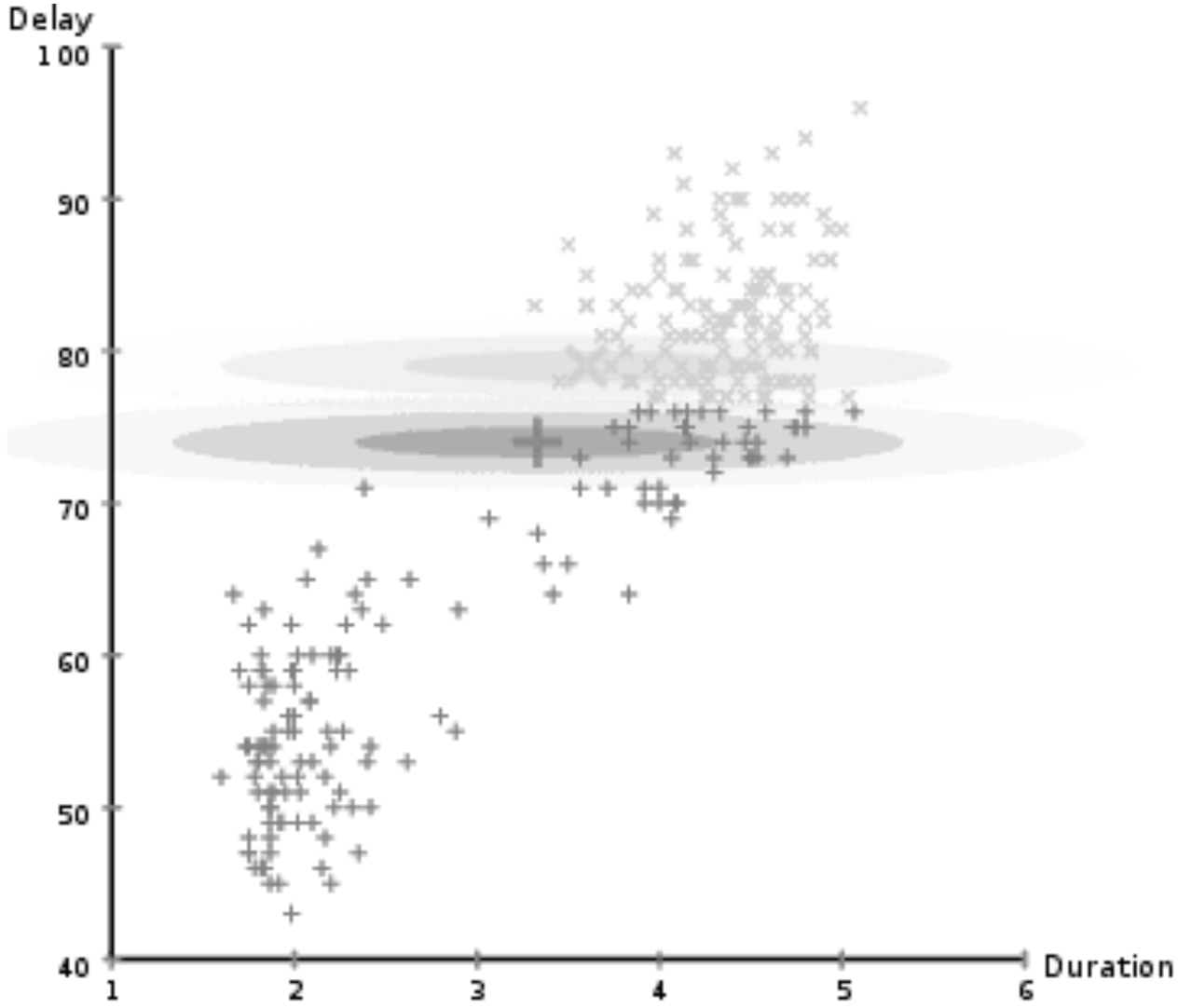
# GMM

Para entrenar el modelo solemos utilizar el algoritmo **expectation maximization** (EM).

Consiste en dos pasos que se aplican iterativamente:

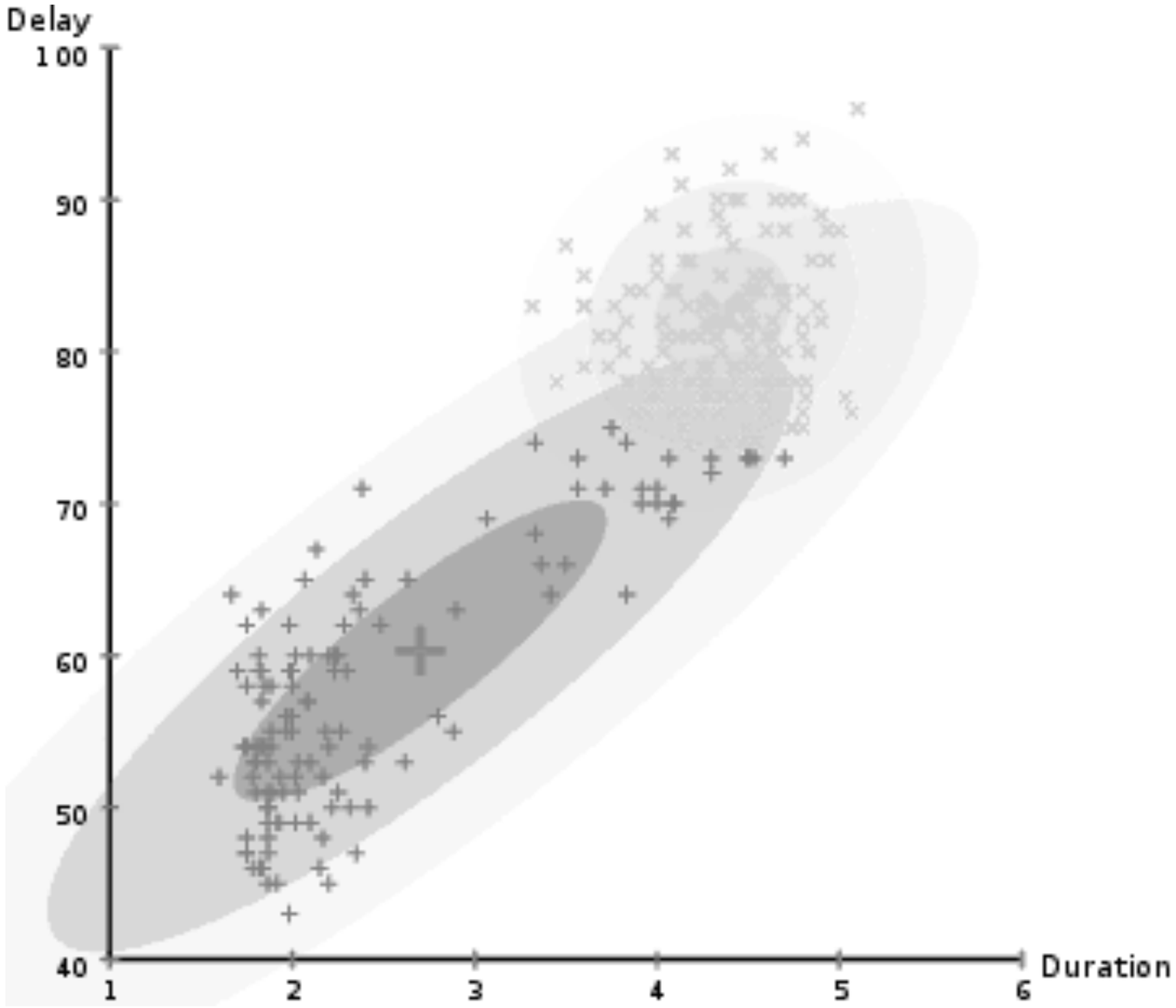
- En el **paso E** asignamos a cada feature vector una de las $N$ gaussianas.
- En el **paso M** optimizamos los parámetros de las $N$ gaussianas.

EM



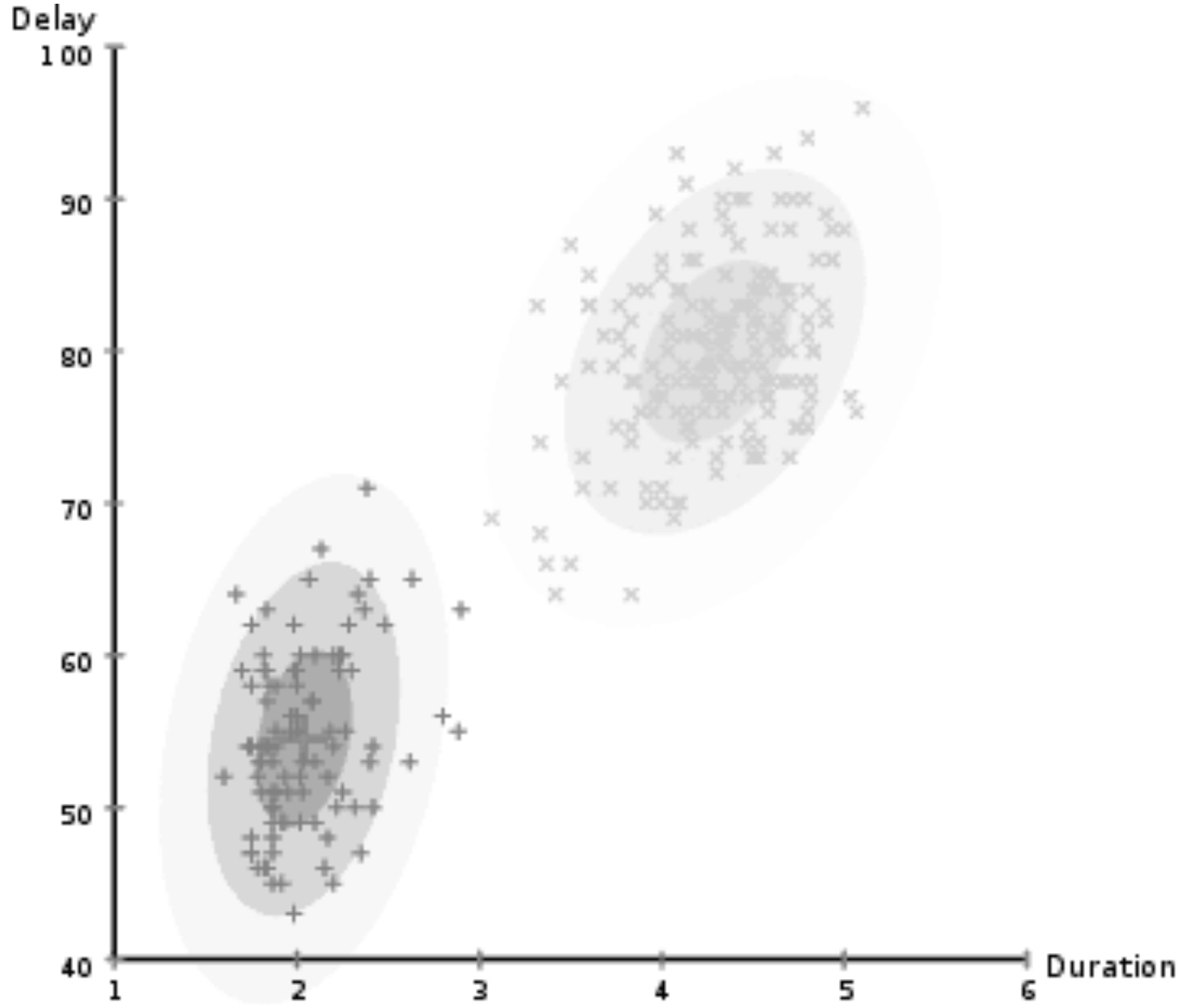

# Algoritmo EM

Paso E




# Algoritmo EM

Paso M




# Algoritmo EM

Resultado final

# Evaluación

Armados con los modelos de locutores $P(X|L)$ y la probabilidad de ocurrencia de locutor $P(L)$, estamos en condiciones de hacer **identificación de locutores 1:N**.

# Evaluación

Para asignar una identidad a una grabación de voz, calculamos los feature vectors $X$ y evaluamos la **verosimilitud logarítmica** de $X$ para cada modelo:

$$\log(P(L \mid X)) = \sum_{t=1}^{T} \log(P(\mathbf{X}_t \mid L)P(L))$$

Los valores resultantes se llaman **scores**.

La identidad del locutor corresponderá al modelo que produzca el máximo score.

# Evaluación

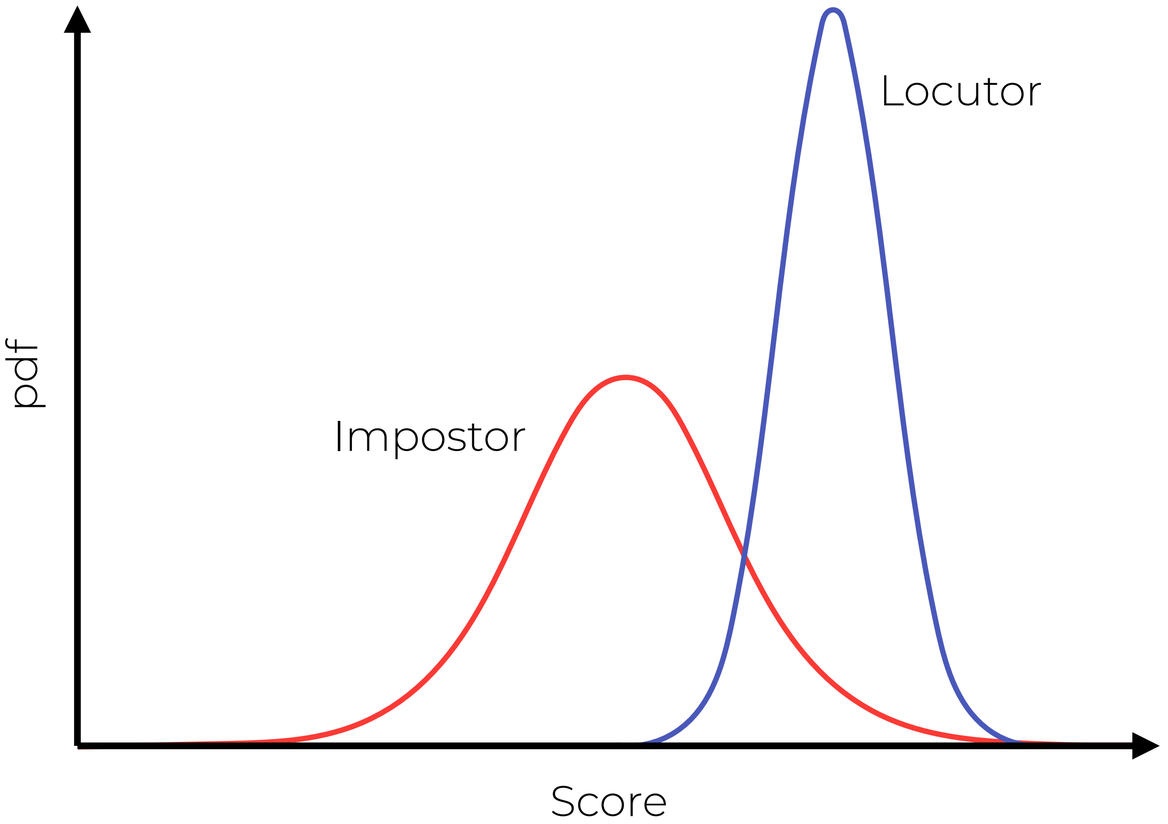
La detección de **locutores desconocidos** requiere otro enfoque, ya que no podemos modelar un locutor desconocido en forma explícita.

Debemos tomar la decisión a partir de un umbral de decisión $t_h$:

$$\log(P(L|X)) \leq t_h$$

# Evaluación

Para determinar $t_h$ estimamos primero la distribución de scores con grabaciones de un locutor, y con grabaciones de un impostor (grabaciones que no son el locutor).



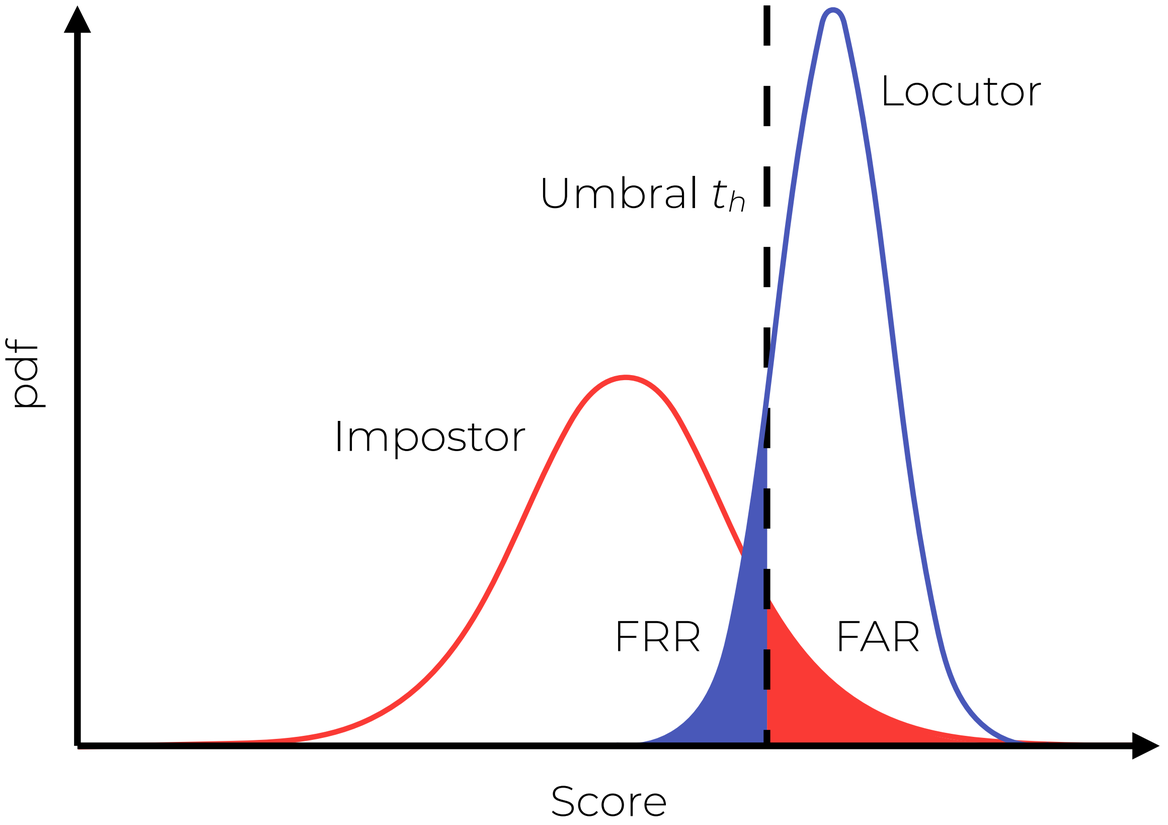
# Evaluación

Dependiendo del valor de $t_h$, tendremos más error por falsa aceptación (FA) o por falso rechazo (FR).

El **false acceptance rate** (FAR) es la probabilidad de aceptación de un impostor (el área roja).

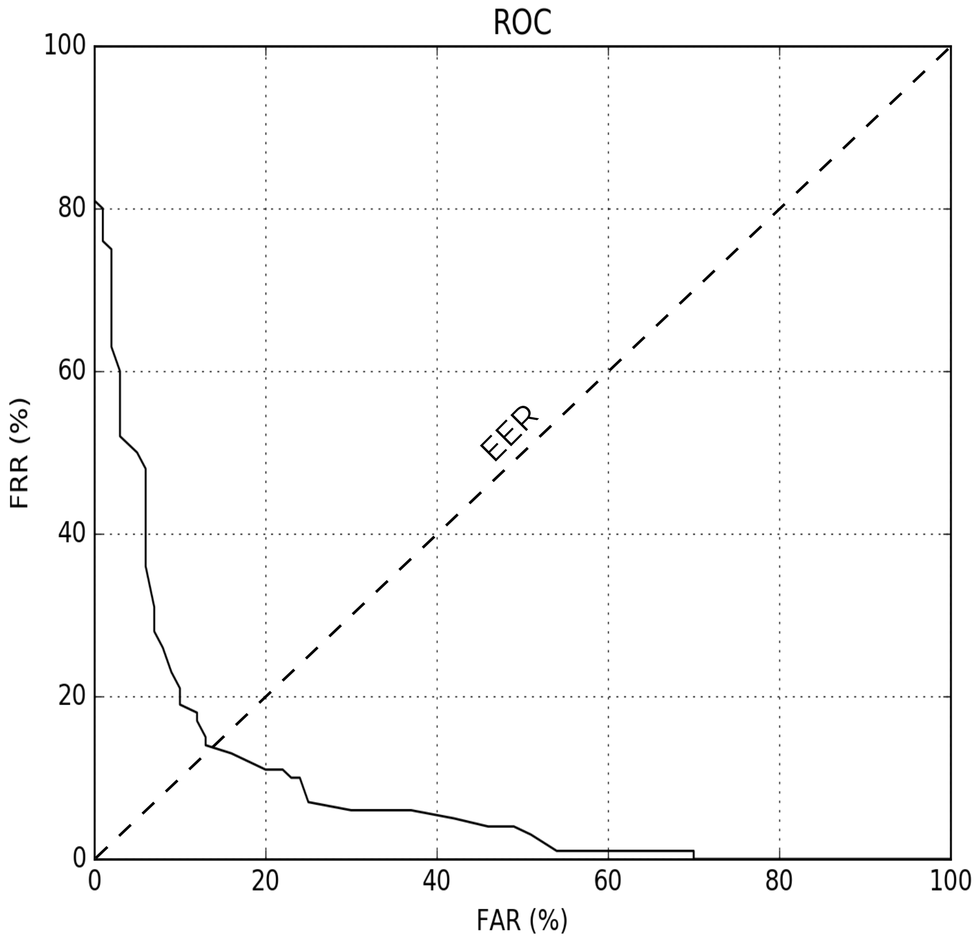
El **false rejection rate** (FRR) es la probabilidad de rechazo del locutor (el área azul).

El **equal error rate** (EER) se da cuando el FAR coincide con el FRR.



# Evaluación

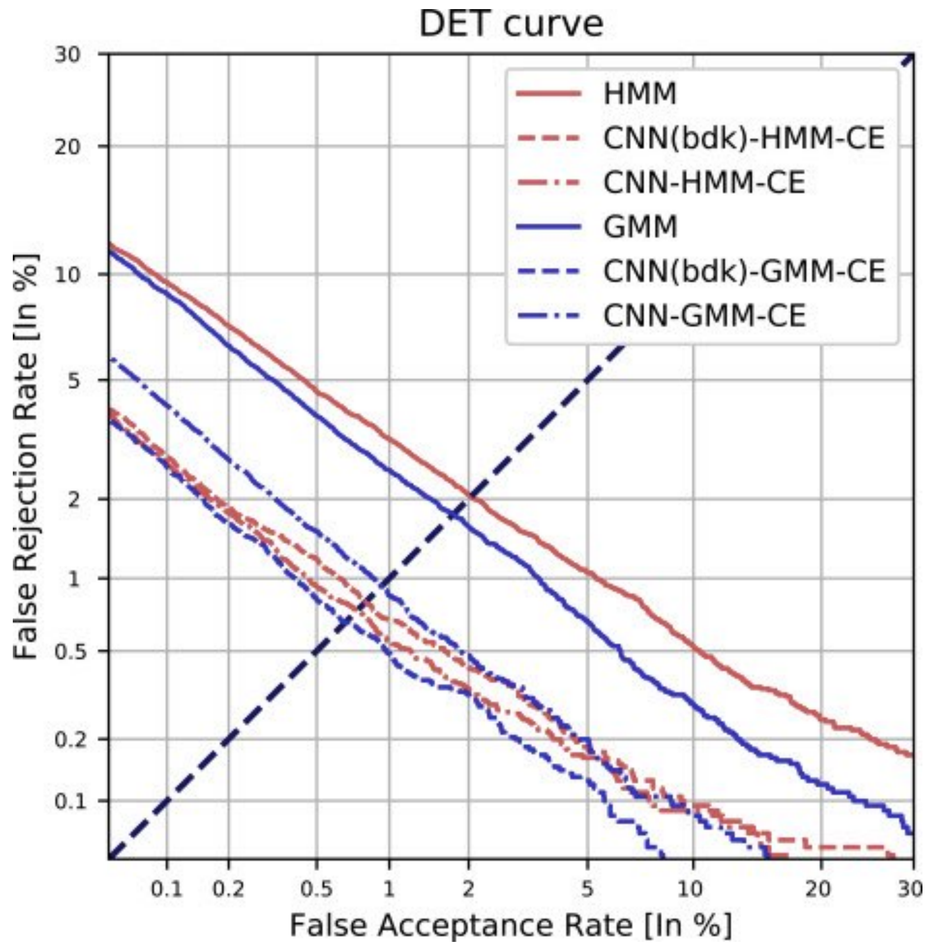
Las **curvas ROC** (receiver operating characteristic) representan el FAR vs. el FRR.



# Evaluación

Las **curvas DET** (detection error trade-off) son curvas ROC representadas en ejes logarítmicos.

Son muy útiles para comparar diferentes sistemas de verificación/identificación de locutores entre sí.



# Evaluación

En **verificación de locutores** aplicamos este procedimiento tal como mostramos.

En **identificación de locutores N+1** determinamos los scores de cada modelo, y si todos los scores son inferiores a cierto umbral $t_h$ que depende de cada modelo, lo declaramos desconocido.

# Enfoque contemporáneo

# Enfoque contemporáneo

Al igual que en los otros problemas que presentamos,  enfoque se basa en **deep learning**.

Suelen utilizarse modelos **end-to-end**.

Requiere un extenso corpus de grabaciones de voz anotadas por locutor.

# Enfoque contemporáneo

Algunos enfoques frecuentes:

- **VAE** con una medida que minimiza la entropía de un mismo locutor y maximiza la entropía entre locutores diferentes.
- **Embeddings**, que modelan las características de cada locutor con un conjunto de valores.