Reconocimiento automático de locutores independiente de texto

Estado del arte y uso en la valoración estadística de la evidencia forense

Daniel Ramos Castro
daniel.ramos@uam.es
ATVS – Biometric Recognition Group
Universidad Autónoma de Madrid
http://atvs.ii.uam.es

Índice y objetivos de la charla

- Sistemas automáticos de reconocimiento de locutor independiente de texto
  - Arquitectura básica
  - Funcionamiento de sistemas basados en información espectral (GMM)
  - Características del estado del arte
    - Resultados ATVS-UAM en NIST SRE 2008

Índice y objetivos de la charla

- Sistemas automáticos de reconocimiento de locutor independiente de texto
  - Arquitectura básica
  - Funcionamiento de sistemas basados en información espectral (GMM)
  - Características del estado del arte
    - Resultados ATVS-UAM en NIST SRE 2008
- Valoración de la evidencia utilizando reconocimiento automático
  - Metodología LR
  - Arquitectura: transformación de score a LR
  - Medida del rendimiento de sistemas basados en LR
    - Importancia y requerimientos
    - Ejemplo: entropía cruzada empírica (ECE)

Índice y objetivos de la charla

- Sistemas automáticos de reconocimiento de locutor independiente de texto
  - Arquitectura básica
  - Funcionamiento de sistemas basados en información espectral (GMM)
  - Características del estado del arte
    - Resultados ATVS-UAM en NIST SRE 2008
- Valoración de la evidencia utilizando reconocimiento automático
  - Metodología LR
  - Arquitectura: transformación de score a LR
  - Medida del rendimiento de sistemas basados en LR
    - Importancia y requerimientos
    - Ejemplo: entropía cruzada empírica (ECE)
- Ejemplo ilustrativo: caso simulado (y muy simplificado)
- Conclusiones

Sistemas automáticos de reconocimiento de locutores independiente de texto

Escuela Politécnica Superior

El problema

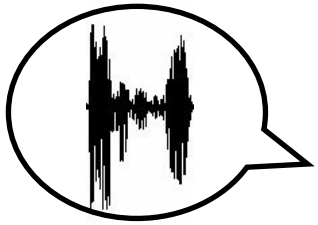
- Grabación incriminatoria (dubitada)
  - Pinchazo telefónico
  - Llamada anónima
  - Micrófono oculto
  - …

Criminal
(Identidad C)

El problema

- Grabación incriminatoria (dubitada)
  - Pinchazo telefónico
  - Llamada anónima
  - Micrófono oculto
  - …
- La policía arresta a un sospechoso
- Se realiza una toma de voz del sospechoso (indubitada)
  - En dependencias policiales
  - Pinchazos cuya autoría se reconoce
  - …

Criminal
(Identidad C)



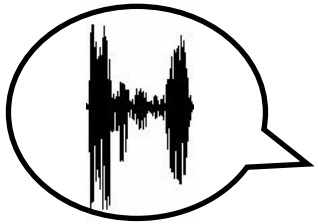
Sospechoso
(Identidad S)

El problema

- Grabación incriminatoria (dubitada)
  - Pinchazo telefónico
  - Llamada anónima
  - Micrófono oculto
  - …

Criminal
(Identidad C)

- La policía arresta a un sospechoso
- Se realiza una toma de voz del sospechoso (indubitada)
  - En dependencias policiales
  - Pinchazos cuya autoría se reconoce
  - …



Sospechoso
(Identidad S)

- El contenido lingüístico no se conoce a priori en ambos casos
  - Independiente de texto

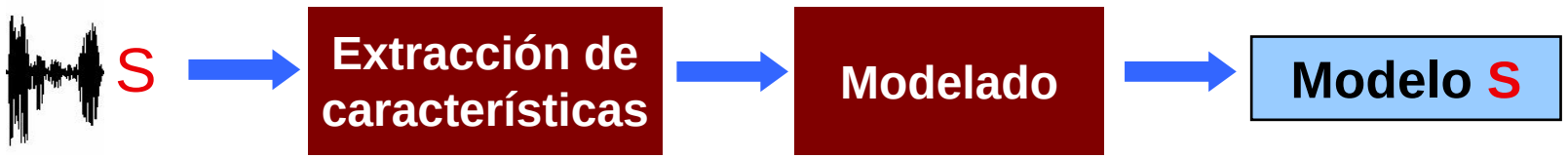
Reconocimiento automático

- La gran mayoría de sistemas calcula puntuaciones (*scores*)
- Similitud entre las identidades en dos fragmentos de voz

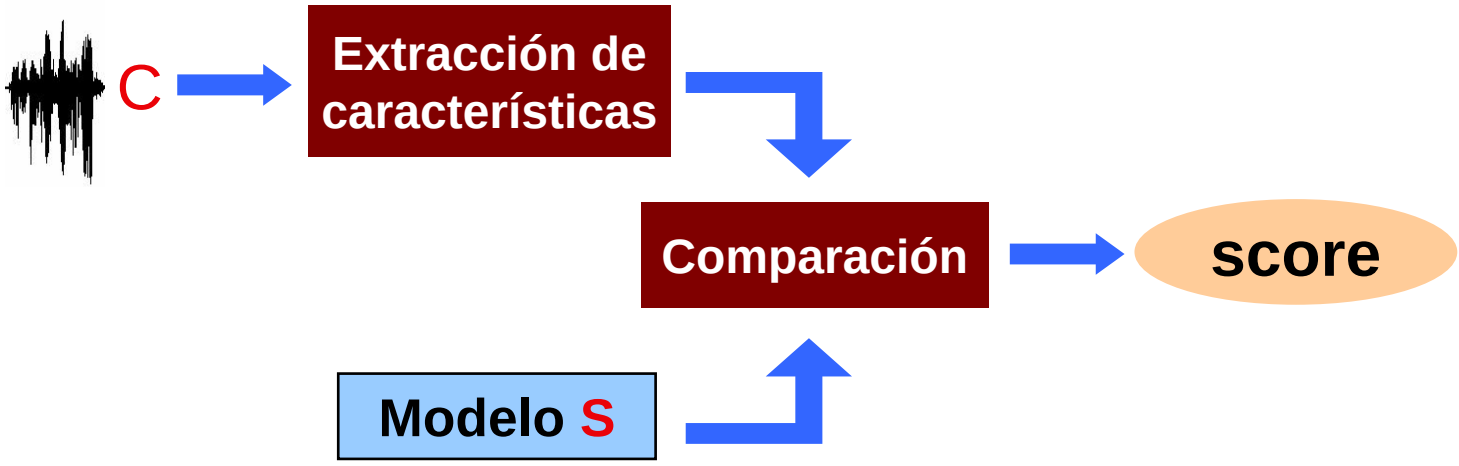
  - Idealmente:
    - Si C y S son la misma identidad, score más alto
    - Si C y S son identidades diferentes, score más bajo
  - Un score permite discriminar

Cálculo de una puntuación (score): etapas

- Modelado de características



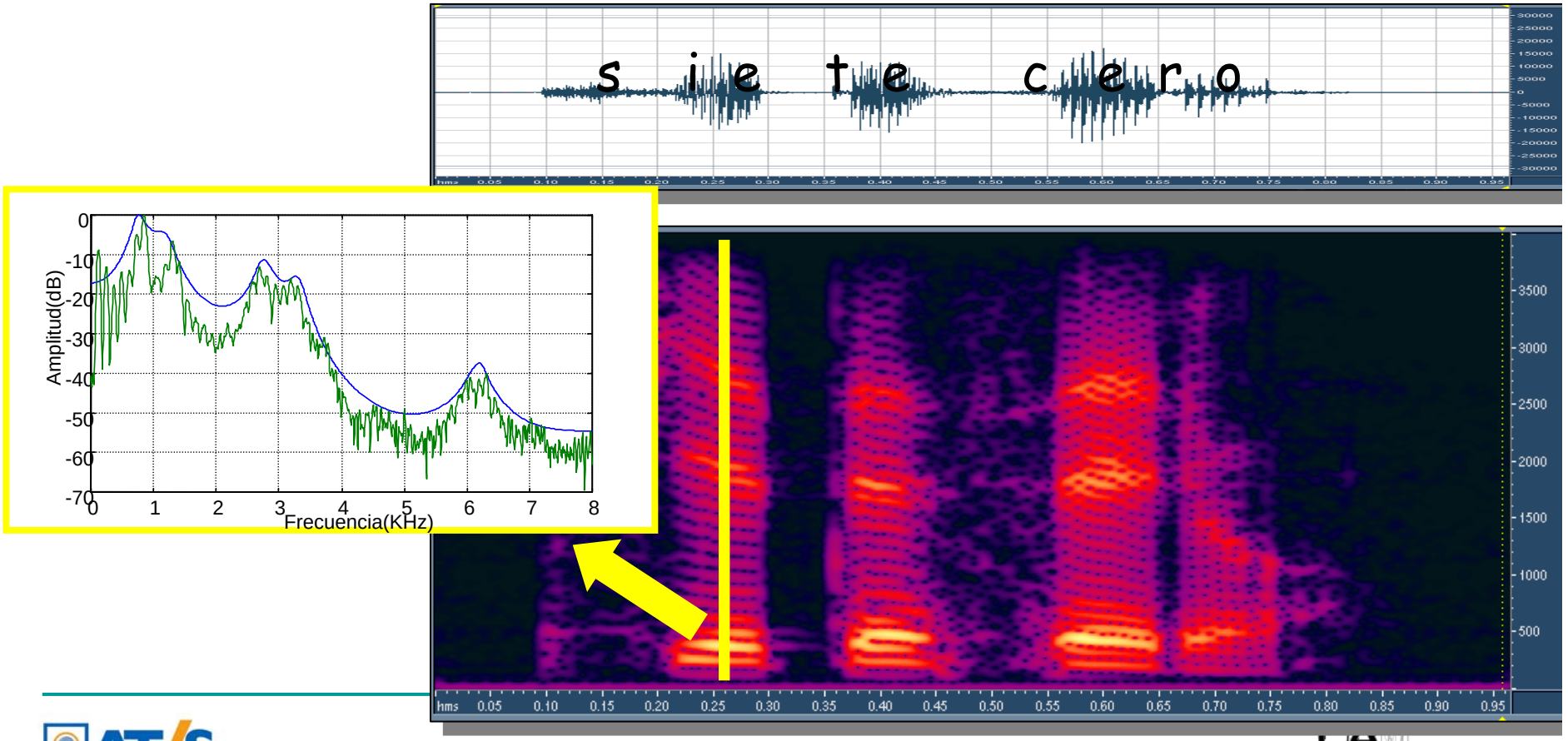
- Cálculo de la puntuación (*score*)



Sistemas basados en información espectral

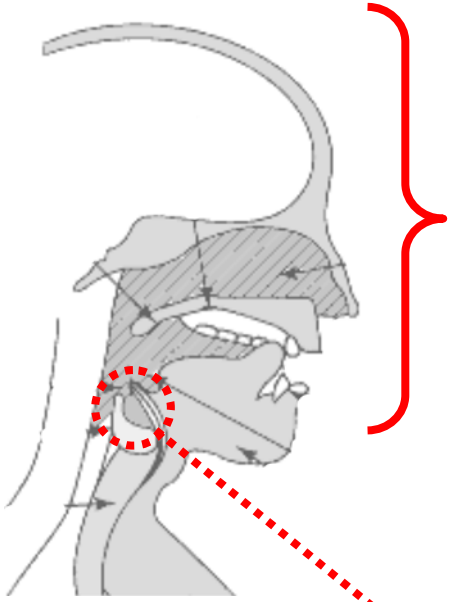
Espectro de la voz

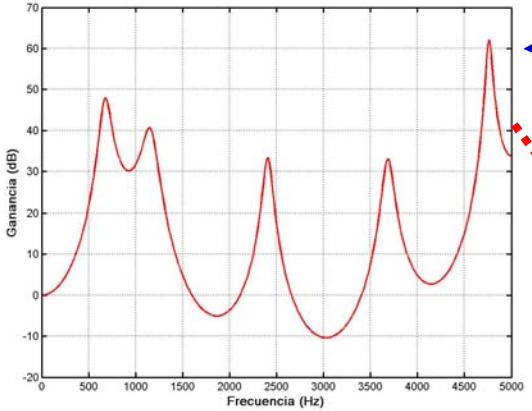
- Sistemas basados en información espectral
  - Extraen información de identidad a partir del espectro y su variación con el tiempo

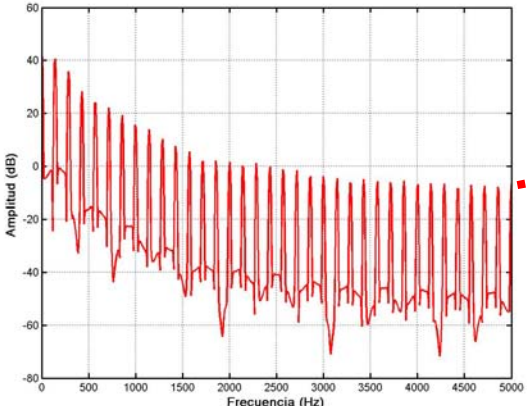


Pulso glotal y resonancia

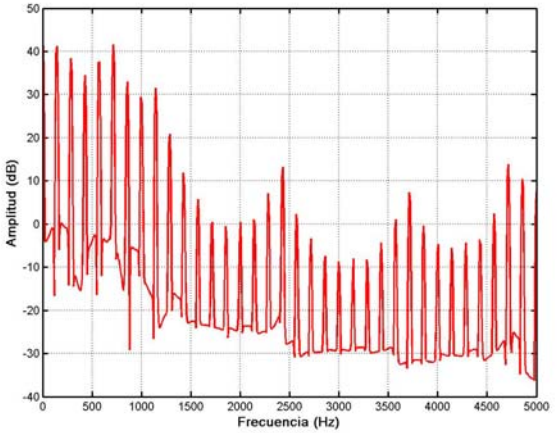
Tracto vocal
(resonancia)







Envolvente espectral: estructura del tracto vocal (particular de cada locutor)



Pulso glotal
(vibración, periódica)

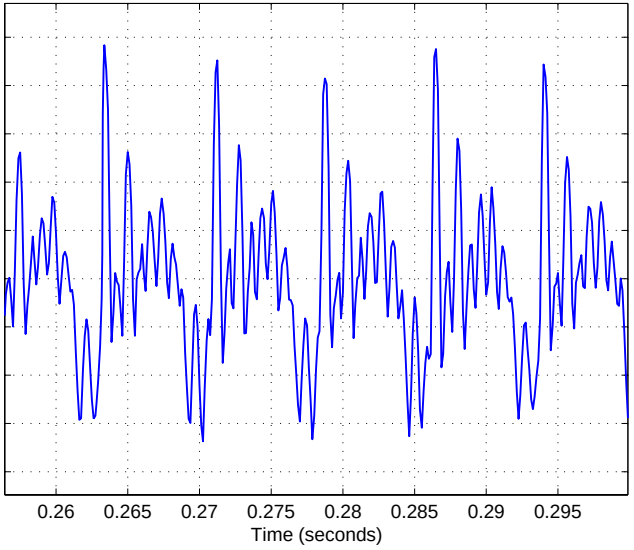
ATVS

UAM UNIVERSIDAD AUTONOMA DE MADRID

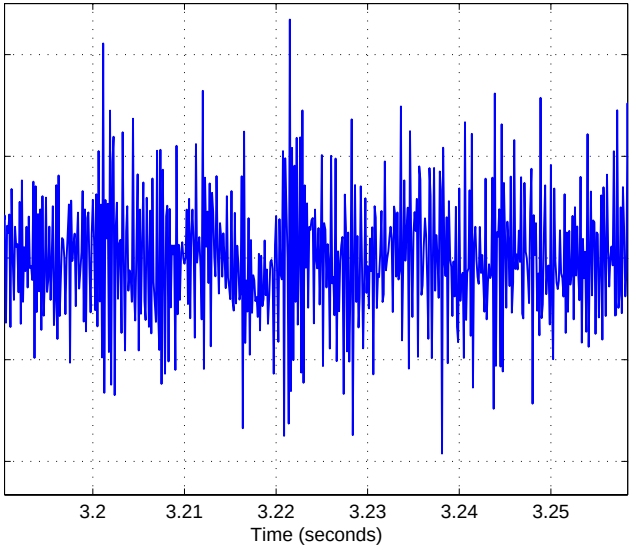
Análisis a corto plazo

- En un periodo corto de tiempo (5-40 ms.) la señal de voz es pseudo-estacionaria
  - Los sonidos sonoros también son pseudo-periódicos

/a/ (sonora)

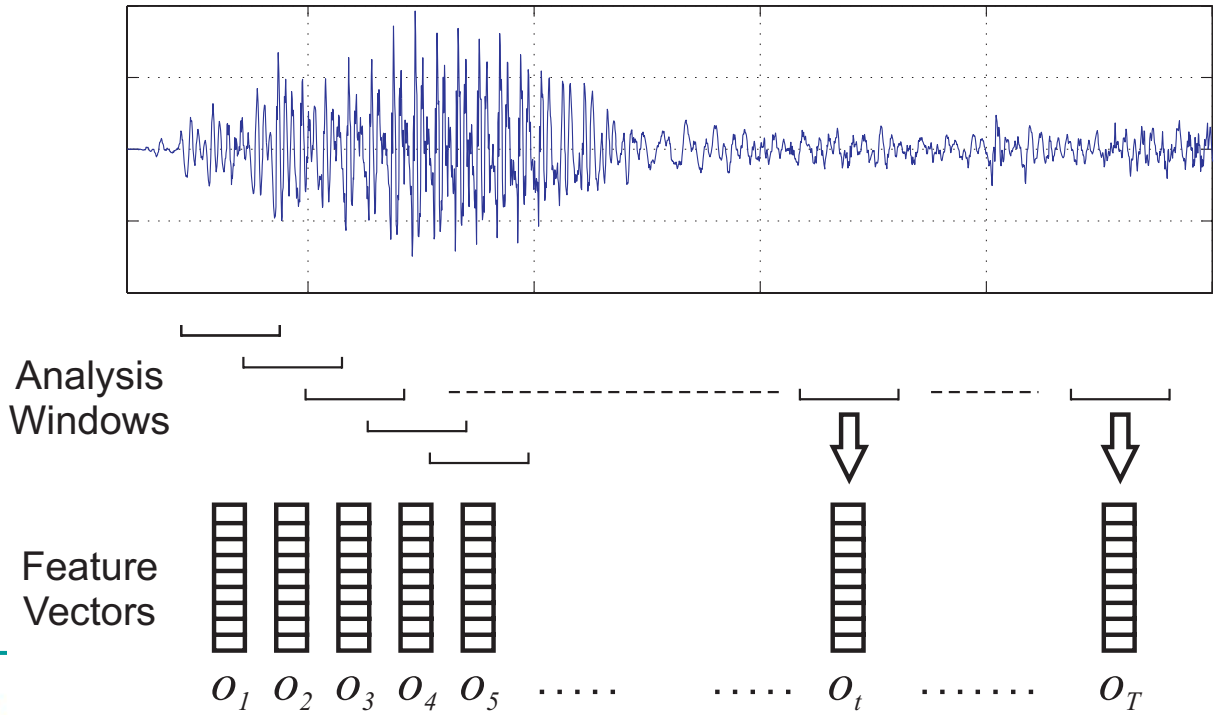


/s/ (sorda)



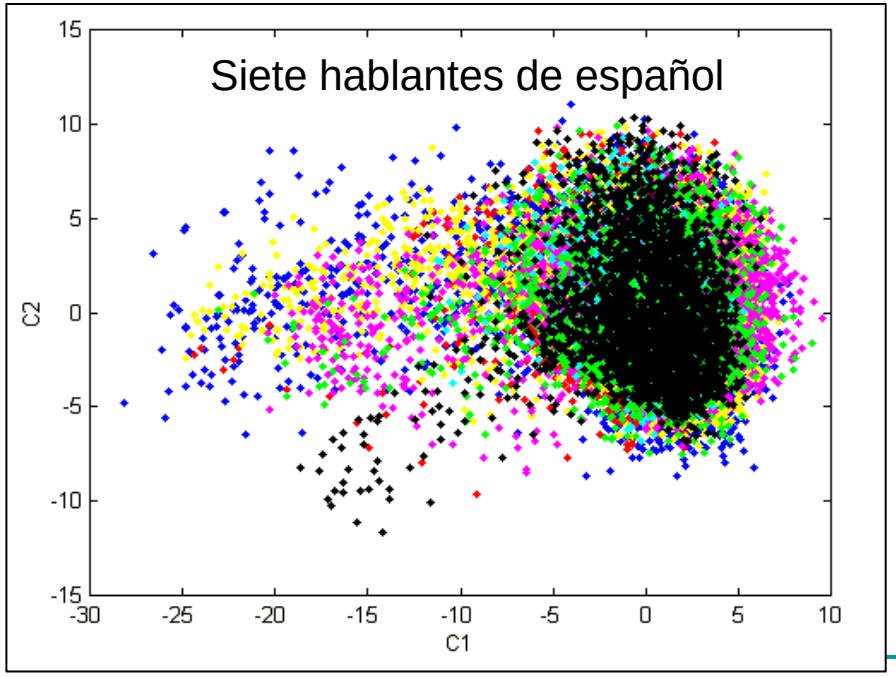
Extracción de características espectrales

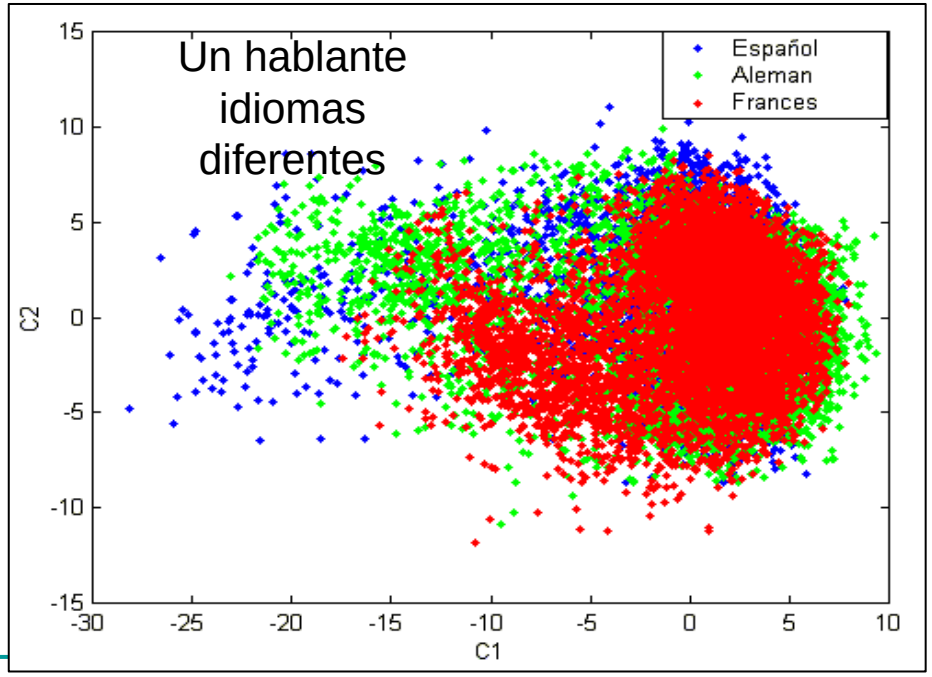
- Primera etapa: enventanado
  - T tramas de voz a corto plazo
- Segunda etapa: extracción de características
  - Cada trama se representa como un vector de D números



Espacio de características espectrales

- Los vectores pertenecen a un espacio vectorial de dimensión D
  - En ese espacio se pueden distinguir vectores de diferentes locutores
- Problemas:
  - Solapamiento entre locutores
  - Variabilidad debida al canal, ruido, idioma, etc.





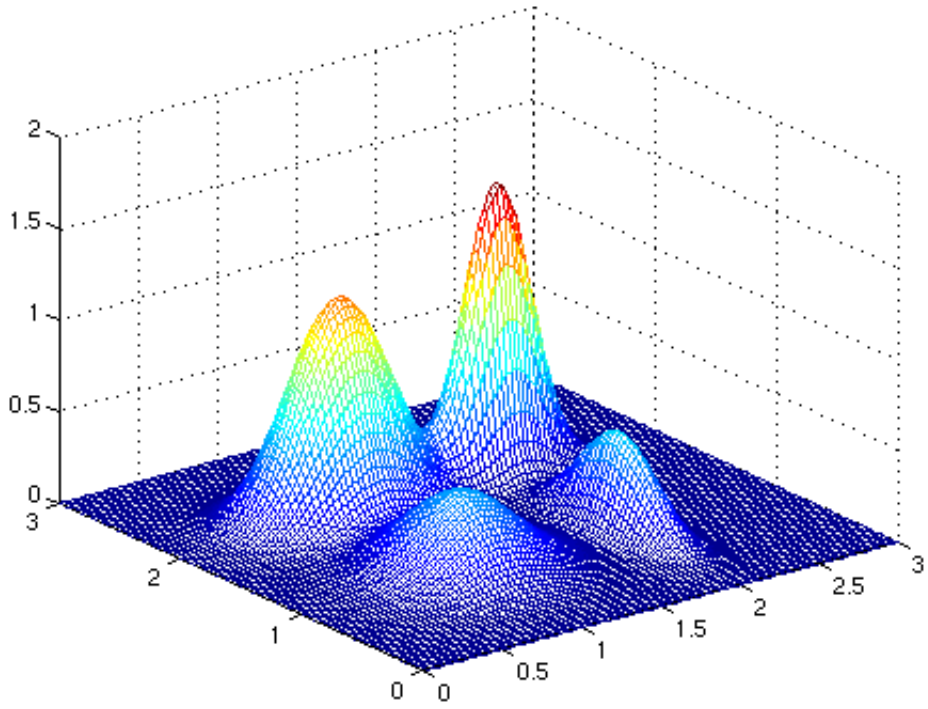
ATVS

UAM UNIVERSIDAD AUTONOMA DE MADRID

Modelos de Mezclas de Gaussianas (GMM)

- Función densidad de probabilidad multidimensional
- Modela la probabilidad de obtener características de un locutor determinado en el espacio
  - Como suma ponderada de densidades de probabilidad gaussianas

- Ejemplo:
  - M=4 componentes (mezclas) gaussianas
  - Espacio de características de D=2 dimensiones
- Detalles en [Reynolds00]



Modelos de Mezclas de Gaussianas (GMM)

Vector de medias (mezcla i): $\mu_p=\{\mu_{ip}\}$

Matriz de covarianzas (mezcla i): $\Sigma_p=\{\Sigma_{ip}\}$

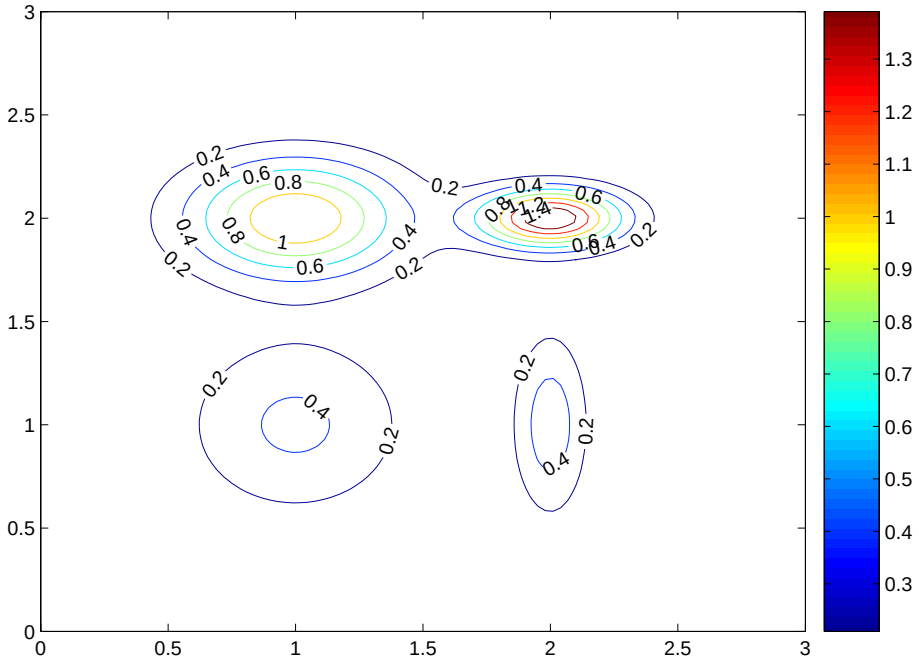
Vector de pesos (mezcla i): $\omega_p=\{\omega_{ip}\}$, $\Sigma_i\ \omega_{ip}=1$

Modelo del locutor p: $\lambda_p=\{\mu_{ip},\Sigma_{ip},\omega_{ip}\}$

$$p\left(\mathbf{o}\middle|\lambda_p\right)=\sum_{i=1}^{M}\omega_{ip}g_{ip}\left(\mathbf{o}\right)$$

$$g_{ip}\left(\mathbf{o}\right)=N\left(\boldsymbol{\mu}_{ip},\Sigma_{ip}\right)$$

- Regiones diferentes del espacio corresponden a configuraciones diferentes del tracto vocal
  - Valores diferentes de las características
- GMM representa bien muy diversas distribuciones de características



Cálculo del score utilizando GMM

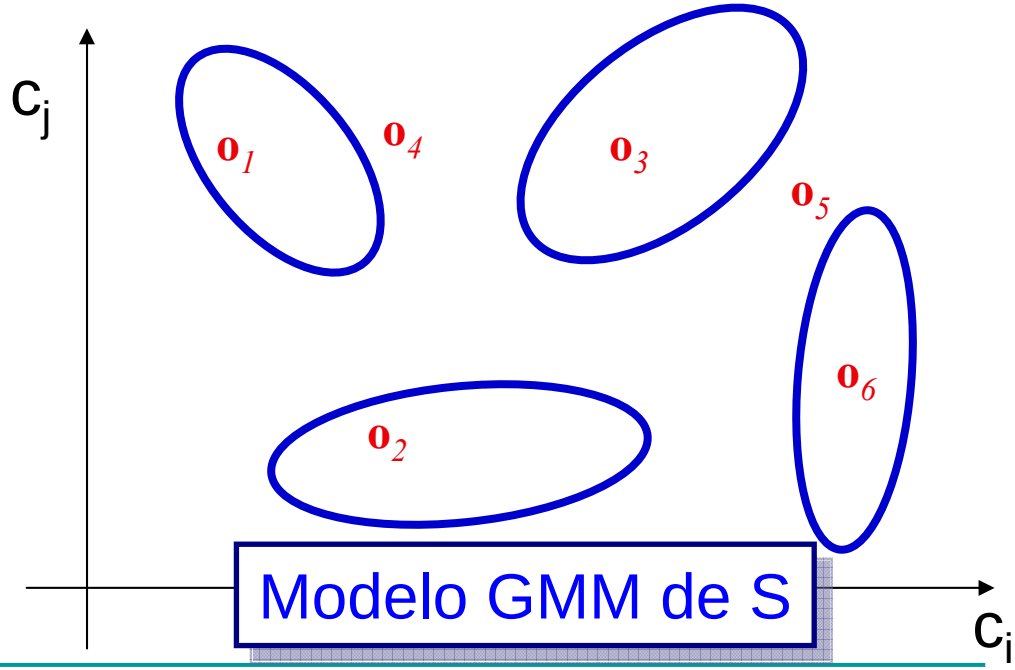
- Partimos del modelo GMM entrenado con el habla de identidad S
- Extraemos características del habla de identidad C
- Cálculo del score:
  - Probabilidad de las muestras suponiendo el modelo
  - Asumiendo independencia entre muestras

C

Extracción de características

$(\mathbf{o}_1,\ldots,\mathbf{o}_6)$

$$p(\mathbf{O}|\lambda_S) = \prod_{t=1}^{T} p\left(\mathbf{o}_t|\lambda_S\right)$$



Otras técnicas de reconocimiento

- Extracción de características
  - Características de alto nivel
    - [Reynolds03,Karajarekar04]
  - Parámetros MLLR
    - [Stolcke06]
- Modelado y cálculo de puntuaciones
  - SVM utilizando supervectores GMM
    - [Campbell06]
  - SVM utilizando kernels GLDS
    - [Campbell06b,Lopez07]
  - HMM (sistemas dependientes de texto)
    - [Rabiner07]
- Normalización de scores [Auckenthaller00]
  - Z-Norm
  - T-Norm

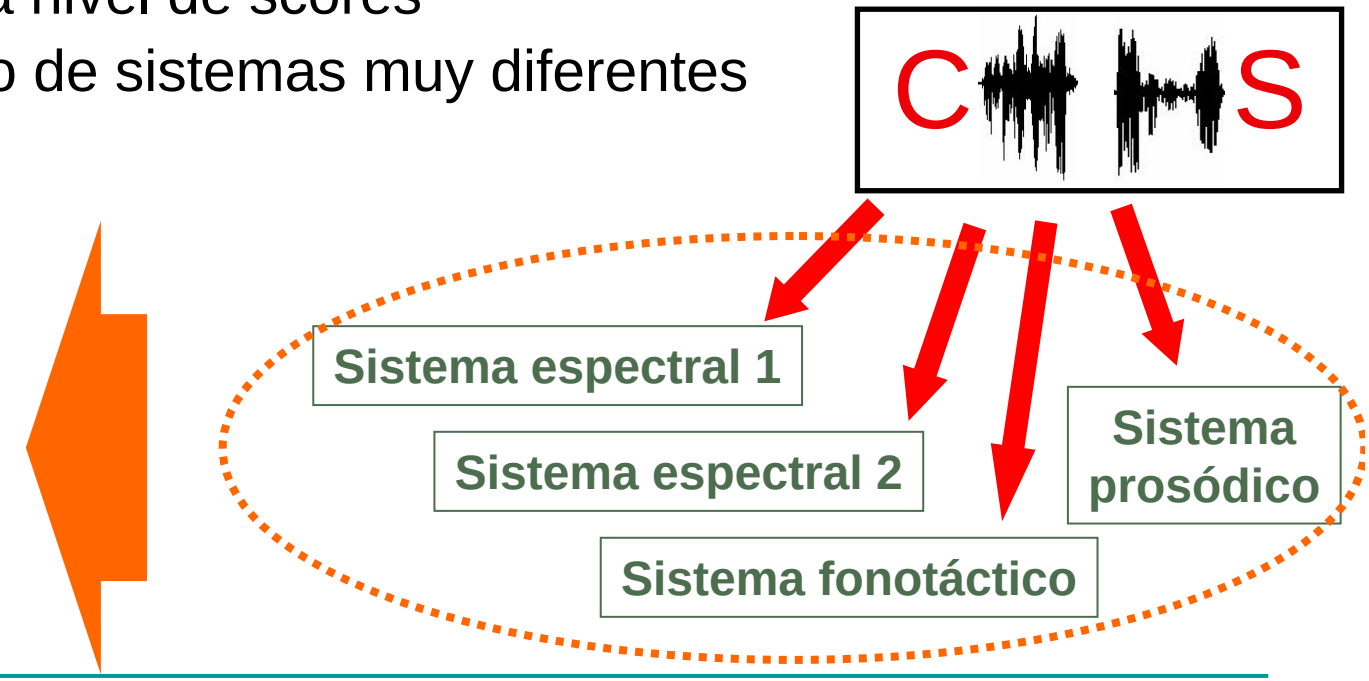
Otras técnicas de reconocimiento

- Extracción de características
  - Características de alto nivel
    - [Reynolds03,Karajarekar04]
  - Parámetros MLLR
    - [Stolcke06]
- Modelado y cálculo de puntuaciones
  - SVM utilizando supervectores GMM
    - [Campbell06]
  - SVM utilizando kernels GLDS
    - [Campbell06b,Lopez07]
  - HMM (sistemas dependientes de texto)
    - [Rabiner07]
- Normalización de scores [Auckenthaller00]
  - Z-Norm
  - T-Norm

Múltiples técnicas diferentes de obtener información sobre la identidad

Fusión de Información

- Muchas técnicas de reconocimiento diferentes
  - Complementarias en muchos casos
  - Se pueden combinar
- Combinación de información: Fusión
  - Típicamente a nivel de scores
  - Permite el uso de sistemas muy diferentes



Desafíos actuales

- Variabilidad de la voz entre sesiones
  - Provoca desajuste en las condiciones del habla dubitada e indubitada (ruido, estilo de habla, reverberación, etc.)
  - [Kenny07,Vogt07]
- Degradación del rendimiento con poco material de voz
  - Locuciones cortas (típicamente dubitadas)
  - [Vogt08,Fauve08]
- Desajuste de base de datos
  - El sistema se entrena con datos en condiciones muy diferentes a la de funcionamiento real (ruido, estilo de habla, reverberación, etc.)
  - [Ramos08]

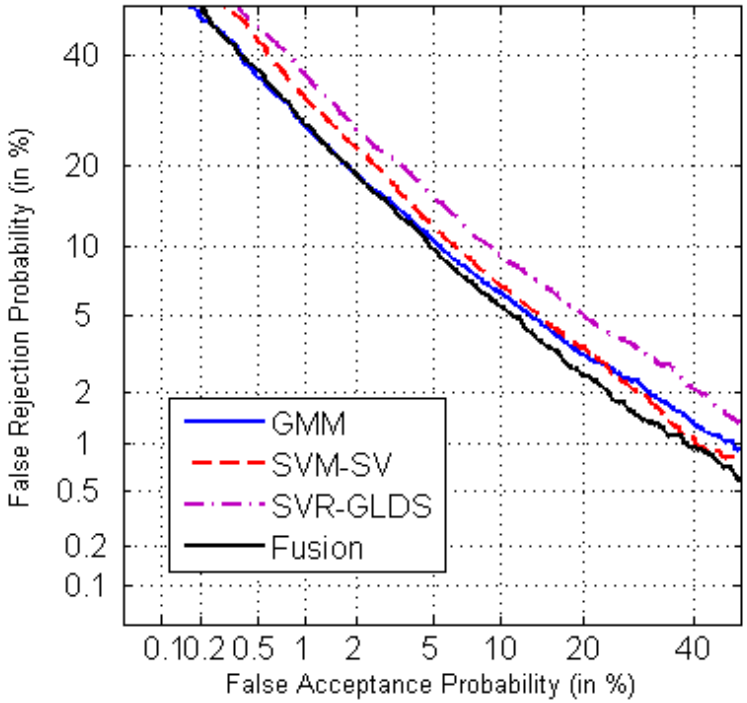
Estado del Arte

- Características (reconocimiento independiente de texto)
  - Dominio de los sistemas espectrales
    - Mucho mejor rendimiento que los sistemas de alto nivel
    - [Reynolds00, Campbell06]
  - Compensación de variabilidad entre sesiones
    - Área de intensa actividad investigadora en la actualidad
    - [Kenny07,Vogt07]
  - Fusión de diferentes sistemas
    - Explotar información complementaria
    - [Brummer07]

Resultados de ATVS-UAM en la evaluación NIST SRE 2008

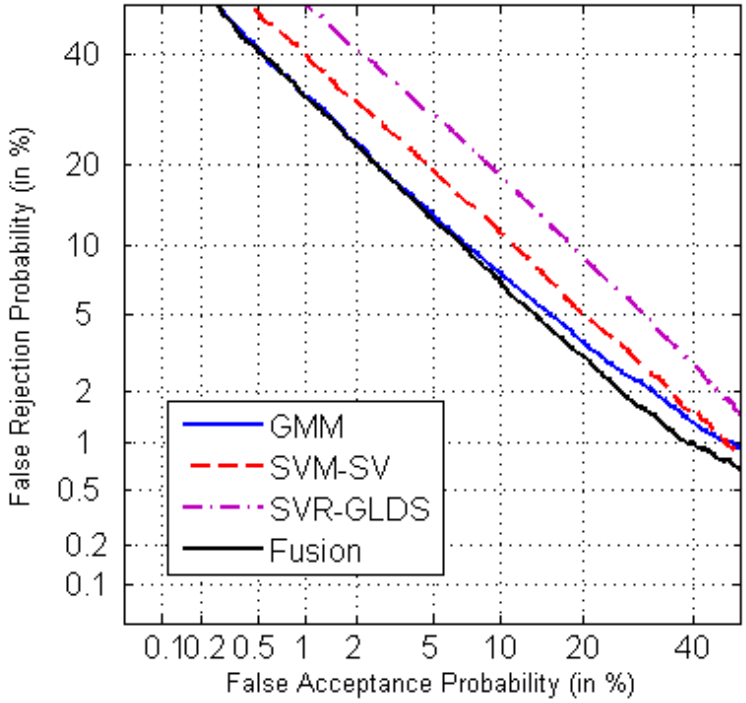
ATVS-UAM en NIST SRE 2008

- Sistema primario ATVS1
  - Fusión de sistemas espectrales con compensación de variabilidad
  - GMM, GMM-SVM-SV, SVM-GLDS
  - Sub-condición teléfono (entrenamiento) vs. teléfono (test)



ATVS-UAM en NIST SRE 2008

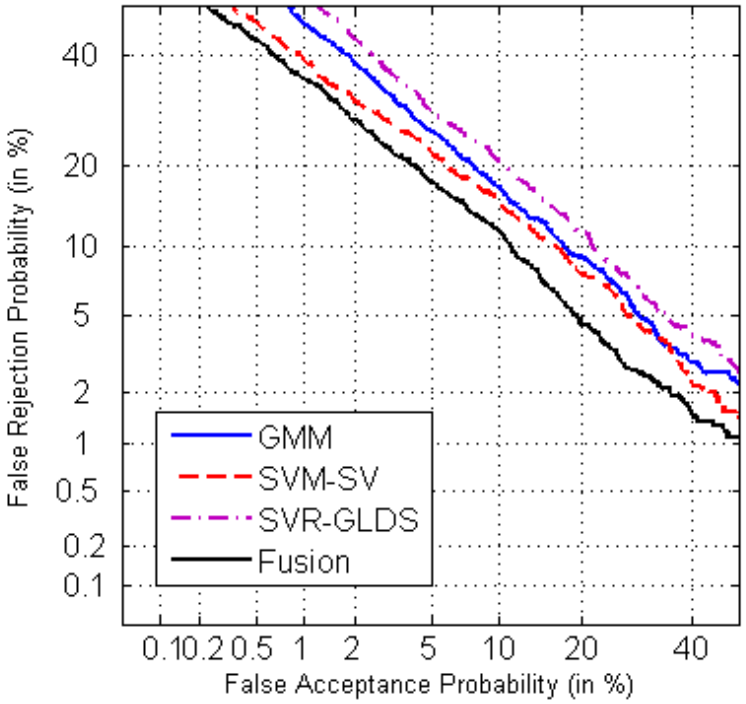
- Sub-condición micrófono (entrenamiento) vs. micrófono (test)
  - 8 diferentes tipos de micrófono, muy diversas calidades
  - Diferentes estilos de habla (conversación, entrevista)



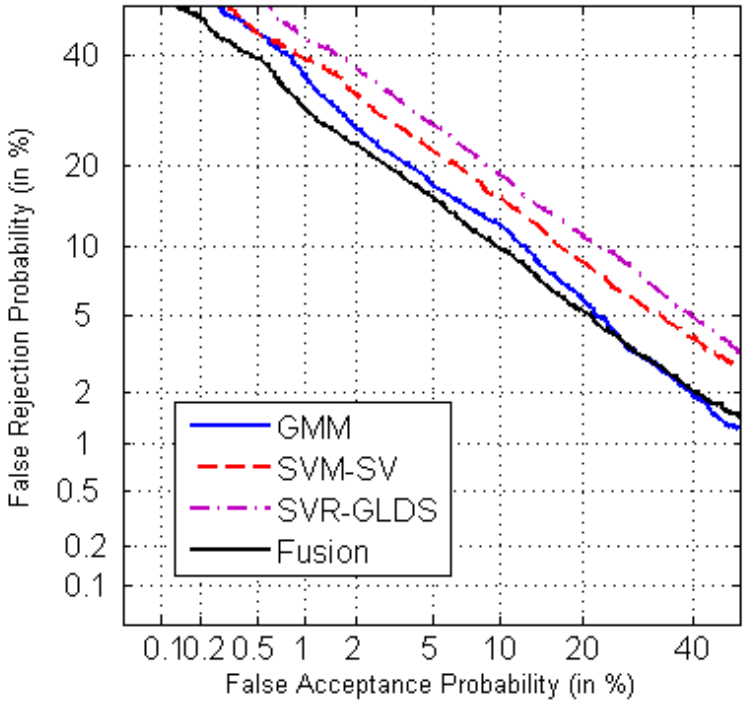
ATVS-UAM en NIST SRE 2008

- Condiciones de desajuste muy fuerte
  - Degradación controlada en condiciones muy adversas
    - Robustez

Teléfono vs. micrófono



Micrófono vs. teléfono

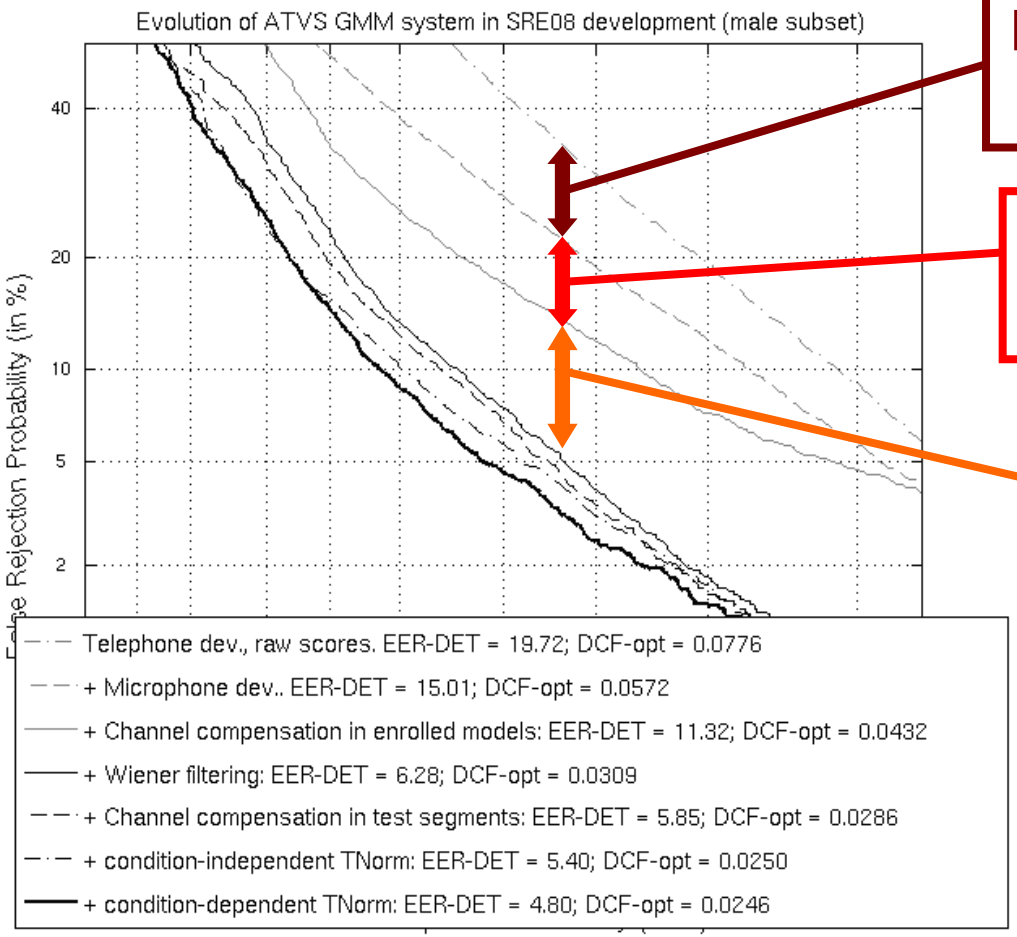


ATVS

UAM UNIVERSIDAD AUTONOMA DE MADRID

Efecto y compensación de variabilidad

- Desarrollo NIST SRE 2008 hombres



Desajuste de base de datos (datos microfónicos)

Compensación de variabilidad entre sesiones

Filtrado adaptativo (filtrado de Wiener)

ATVS

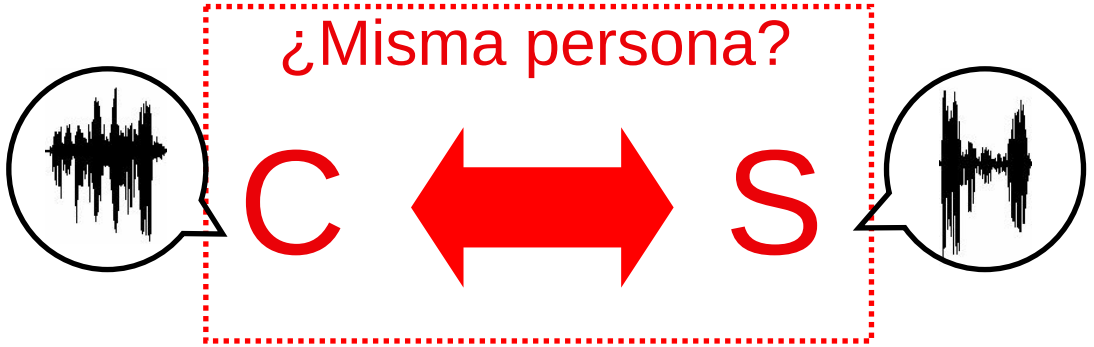
UAM UNIVERSIDAD AUTONOMA DE MADRID

Valoración de la evidencia forense utilizando sistemas automáticos de reconocimiento de locutor

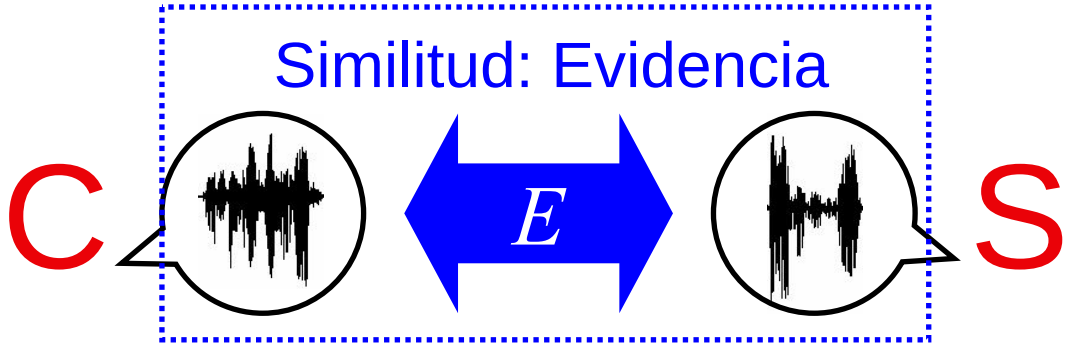
Escuela Politécnica Superior

Preguntas y evidencias

- Lo que quiere saber el juez: ¿pertenecen la toma dubitada y la indubitada a la misma persona?

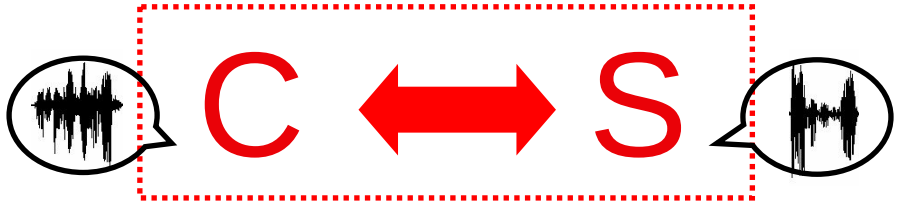


- Sistema automático (perito): mide similitud entre voces



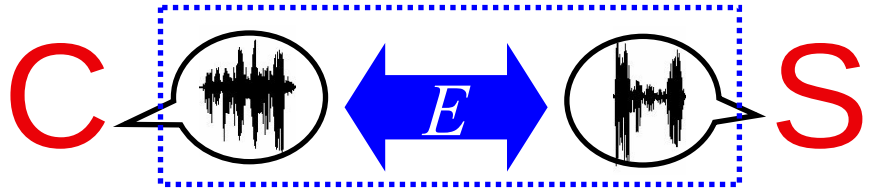
Metodología LR

- Lo que quiere saber el juez: ¿son la misma persona?



$$\frac{P\left(\theta_p \middle| E, I\right)}{P\left(\theta_d \middle| E, I\right)}$$

- Trabajo del perito: valoración de la evidencia



$$LR = \frac{p\left(E \middle| \theta_p, I\right)}{p\left(E \middle| \theta_d, I\right)}$$

- Peso de la evidencia: relación de verosimilitudes (LR)
  - Cociente de probabilidades de observar la similitud dubitada-indubitada suponiendo
    - θ_p: el sospechoso es el autor de la toma dubitada
    - θ_d: otro individuo en la población es el autor de la toma dubitada

Separación de roles

Perito Forense

BBDD

Científico forense

$$LR = \frac{p\left(E\middle|\theta_p, I\right)}{p\left(E\middle|\theta_d, I\right)}$$

LR=1000

Inferencia de la identidad

$$\frac{P\left(\theta_p\middle|I\right)}{P\left(\theta_d\middle|I\right)} = \frac{1}{99}$$

$$\frac{P\left(\theta_p\middle|E,I\right)}{P\left(\theta_d\middle|E,I\right)} = \frac{1000}{99}$$

LR *apoya* opinión previa
(no es opinión en sí)

ATVS

UAM UNIVERSIDAD AUTONOMA DE MADRID

Cálculo del LR utilizando sistemas automáticos

- El sistema automático genera un score
  - Arquitectura básica del sistema: en general no modificable
- Score mide similitud entre identidades: evidencia E
  - No necesariamente interpretable como un valor de LR
- Paso necesario: transformar el score en un LR

Reconcimiento automático de locutor

Transformación score a LR

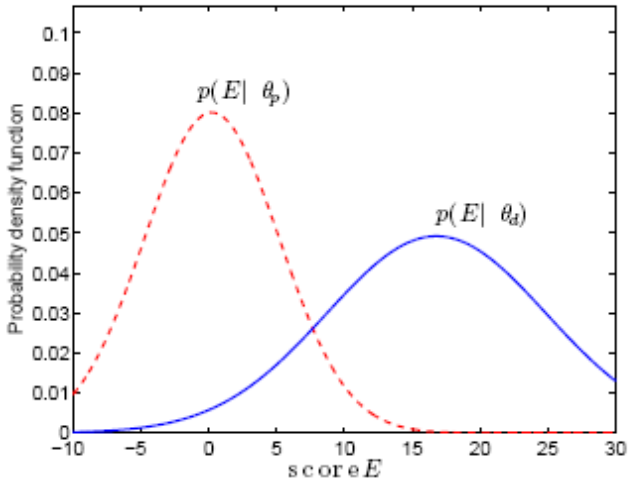
LR

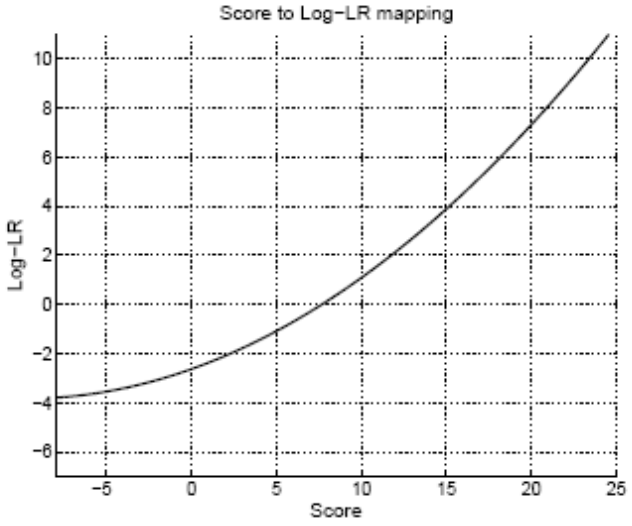
Score: evidencia (E)

$$LR = \frac{p\left(E \middle| \theta_p, I\right)}{p\left(E \middle| \theta_d, I\right)}$$

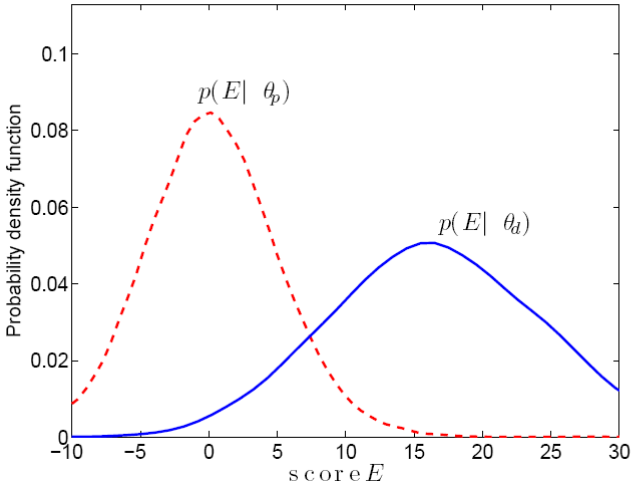
Transformación de score a LR

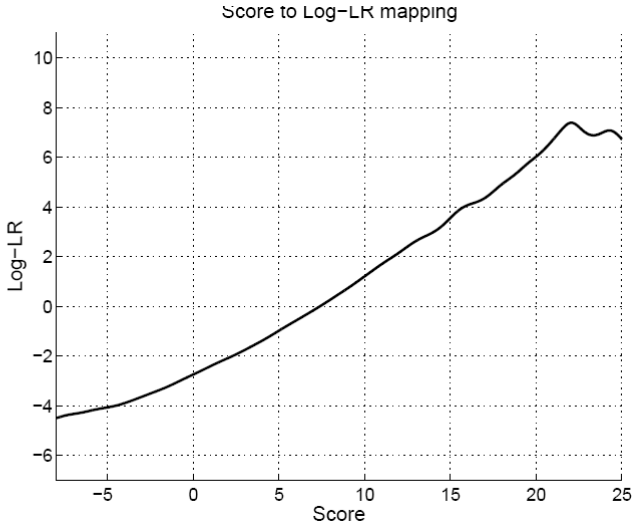
- Técnicas generativas [Meuwly00,Ramos07]:





Modelado gaussiano



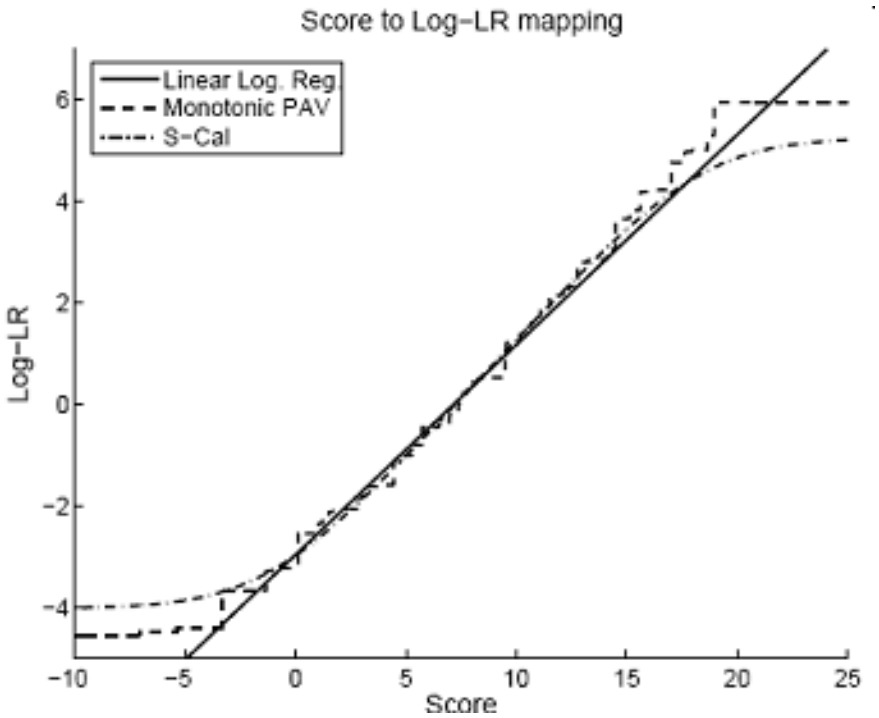


Kernel Density (KDF)

UAM UNIVERSIDAD AUTONOMA DE MADRID

Transformación de score a LR

- Otras técnicas [Brummer07]:



  - Regresión logística (lineal)
  - S-cal
  - Pool Adjacent Violators (PAV)

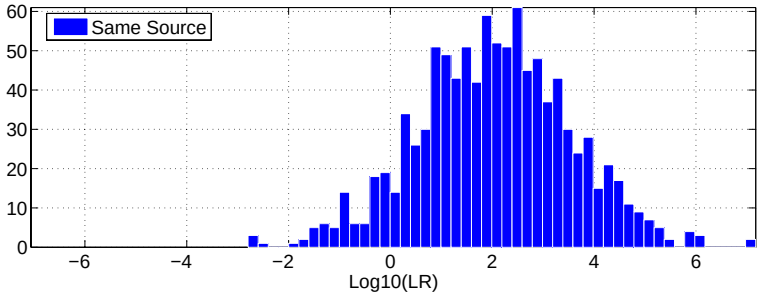
Medida del rendimiento de sistemas basados en LR

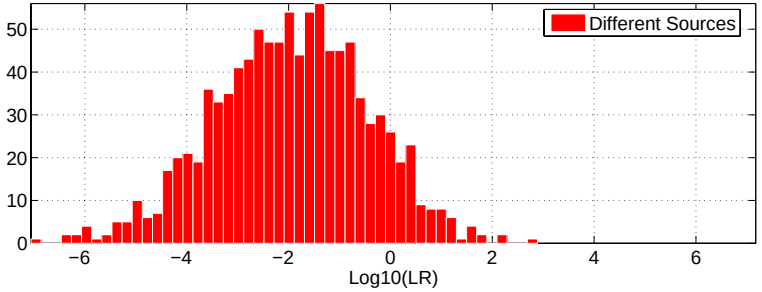
Medida del rendimiento: importancia

- En ciencia forense, se demanda cada vez más un enfoque científico de medida del rendimiento
  - Repetibilidad
  - Prueba empírica
- Varias razones
  - Errores de identificación en disciplinas supuestamente infalibles
    - Caso Mayfield en huella dactilar
  - Evolución del análisis de ADN
    - Metodología científica, basada en datos, probabilística, aceptada
  - Impulso/impacto de las reglas Daubert americanas
    - Demandan procedimientos de medida del rendimiento potencial o real de las técnicas

Medida empírica del LR

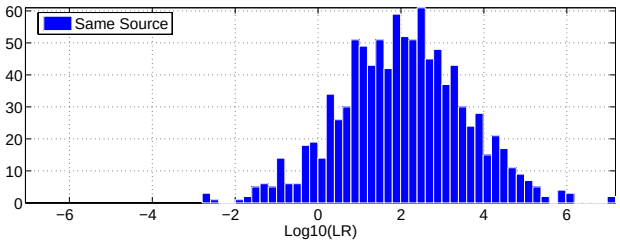
- Prueba experimental
  - Base de datos de ficheros de voz
    - Condiciones lo más parecidas posible a las de un caso dado
    - Las identidades de los fragmentos de habla son conocidas
  - Generar LR en los que θ_p es cierta (misma fuente)
  - Generar LR en los que θ_d es cierta (fuentes diferentes)

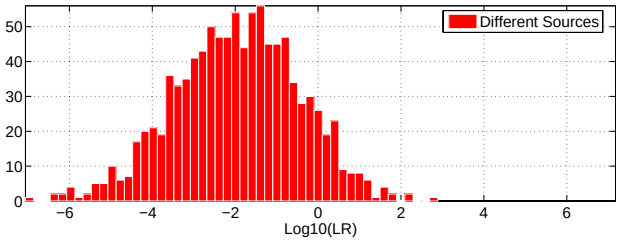


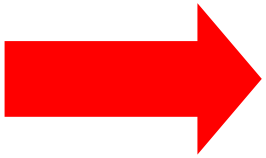


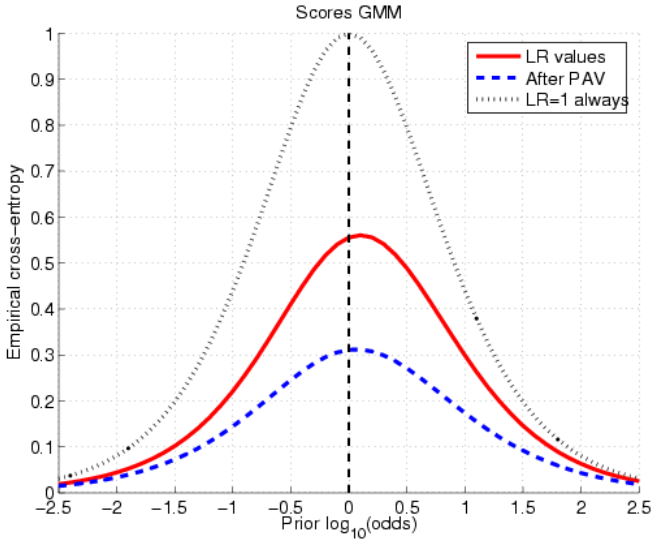
- Metodología científica
  - Repetible
  - Basada en datos
  - …

Ejemplo de medida del rendimiento (precisión): Entropía cruzada empírica (*ECE*)







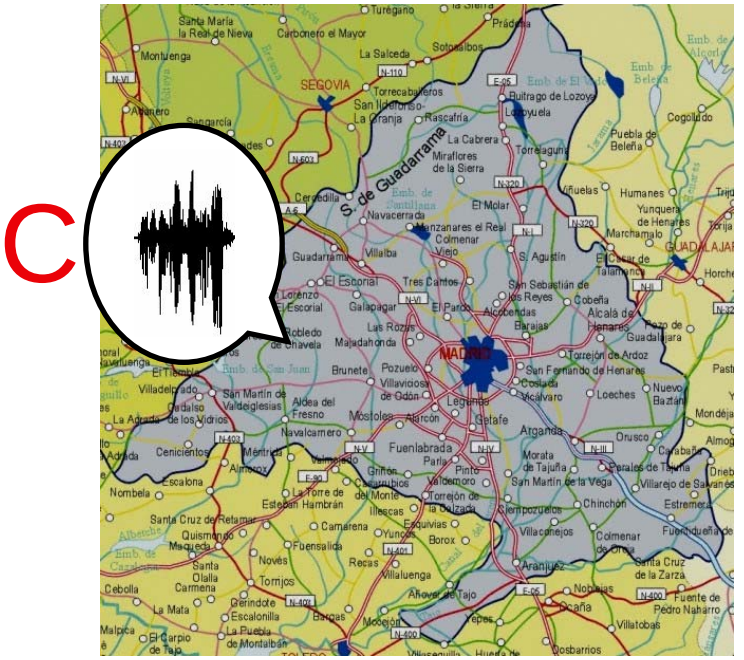


- "Información media para obtener certeza"
  - Cuanto más alta *ECE*, peor es la técnica analizada
- Ventaja: separación de roles:
  - Perito: puede calcular *ECE* a partir únicamente del conjunto experimental del LR
    - Asumiendo un rango amplio de otra información (probabilidades) a priori
  - Juez: con el resto de información (probabilidad) a priori podría calcular *ECE*
- Detalles en [Ramos07]

Ejemplo ilustrativo: caso simulado (y muy simplificado)

Caso simulado

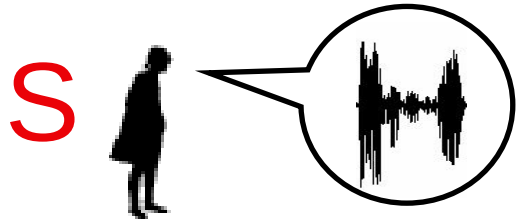
- Grabaciones incriminatorias tomadas en la Comunidad Autónoma de Madrid
  - Población: potenciales criminales
    - Hablantes de Madrid con características similares al hablante de la toma dubitada
      - Idioma
      - Acento
      - ...
  - Sistema: GSM grabado en cinta magnetofónica



- Las investigaciones policiales llevan a la detención de un sospechoso

Caso simulado

- Se realizan grabaciones del sospechoso (voz indubitada)



  - En principio, la abundancia y control sobre las grabaciones suele ser mayor que en la toma dubitada
    - Pero posiblemente en condiciones muy diferentes a la toma dubitada
  - Puede haber incluso pinchazos no incriminatorios de los cuales el sospechoso reconoce la autoría
    - Condiciones similares a la toma dubitada

- El juez le pide al perito:
  - Que evalúe la evidencia
  - Que le informe de la precisión de las técnicas utilizadas

Cálculo del LR

- Paso 1: el sistema automático calcula un score
  - Sin valor por sí mismo
    - ¿10 con respecto a qué?
  - En general, no interpretable
    - A priori, no conocemos su rango de variación

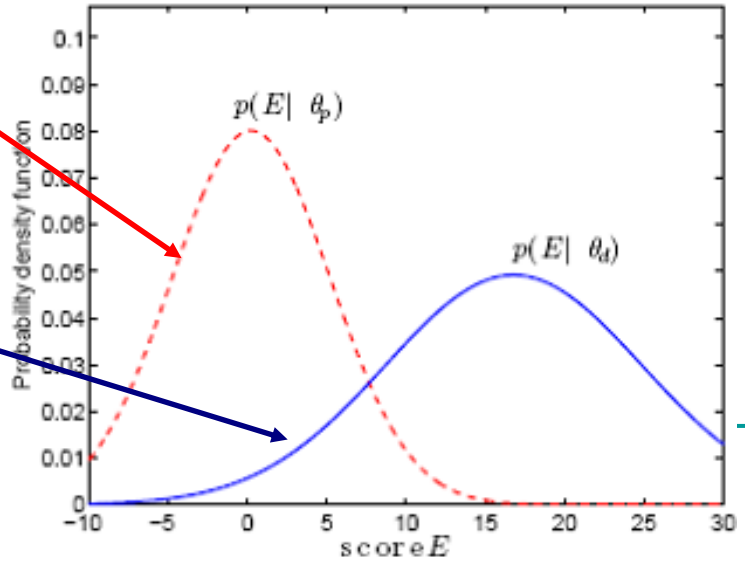
Cálculo del LR

- Paso 1: el sistema automático calcula un score
  - Sin valor por sí mismo
    - ¿10 con respecto a qué?
  - En general, no interpretable
    - A priori, no conocemos su rango de variación

- Paso 2: cálculo del LR
  - En este ejemplo usamos modelado gaussiano

Intervariabilidad (población)

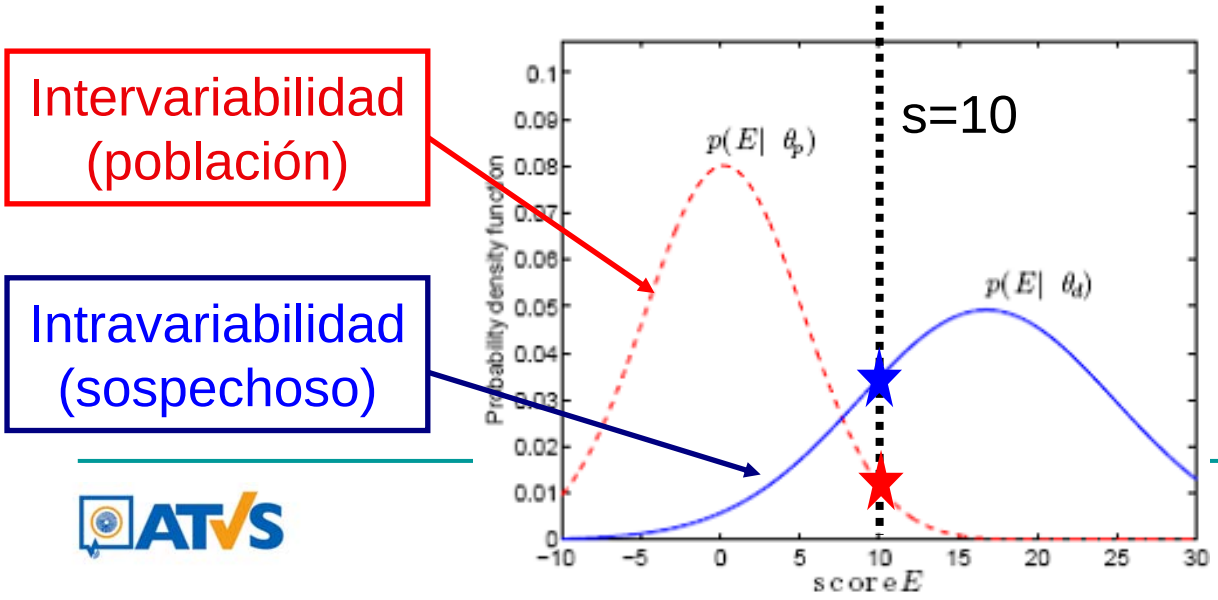
Intravariabilidad (sospechoso)



Cálculo del LR

- Paso 1: el sistema automático calcula un score
  - Sin valor por sí mismo
    - ¿10 con respecto a qué?
  - En general, no interpretable
    - A priori, no conocemos su rango de variación

- Paso 2: cálculo del LR
  - En este ejemplo usamos modelado gaussiano



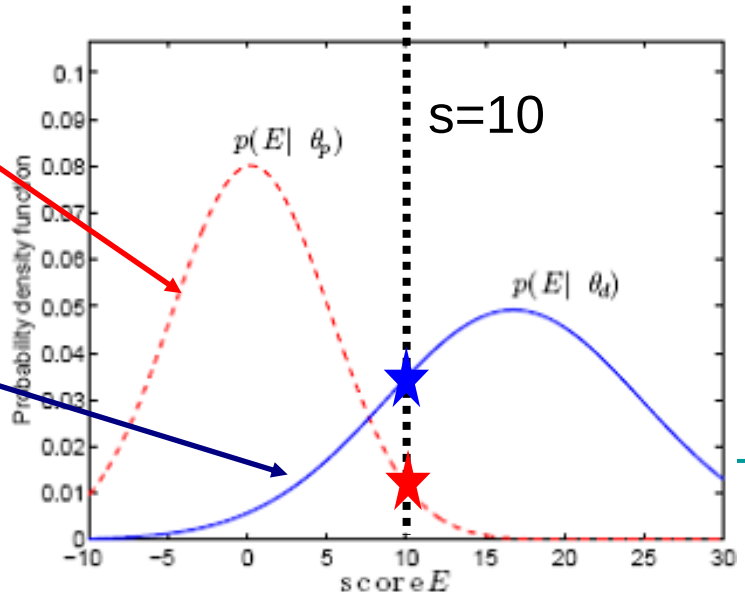
Cálculo del LR

- Paso 1: el sistema automático calcula un score
  - Sin valor por sí mismo
    - ¿10 con respecto a qué?
  - En general, no interpretable
    - A priori, no conocemos su rango de variación

- Paso 2: cálculo del LR
  - En este ejemplo usamos modelado gaussiano

Intervariabilidad (población)

Intravariabilidad (sospechoso)



$$LR = \frac{0,35}{0,15} = 2,33$$

Apoyo 2,33 a 1 a la hipótesis θ_p (“misma fuente”)

ATVS

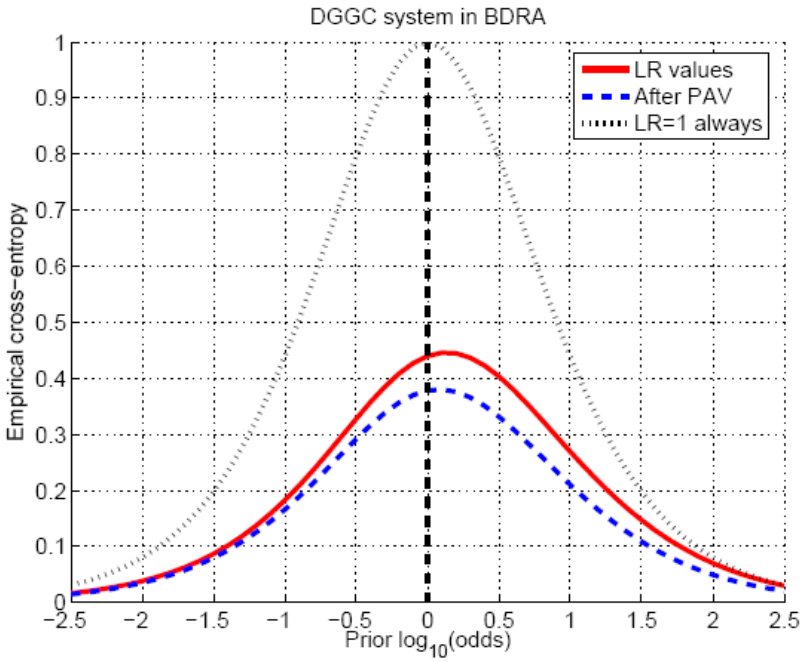
UAM UNIVERSIDAD AUTONOMA DE MADRID

Medida del rendimiento

- El perito debe evaluar la precisión de su sistema
  - Para ello, utiliza un enfoque empírico y repetible
- Para este ejemplo:
  - Sistema utilizado en el Departamento de Acústica e Imagen (DAI) de la Guardia Civil
  - Base de datos Ahumada III [Ramos08]:
    - Subcorpus de BDRA (Base de Datos de Registros Acústicos)
      - Orden ministerial INT-3764-04, BOE 277
    - Habla proveniente de casos forenses reales autorizados
    - Habla GSM grabada sobre cinta magnetofónica
      - Condiciones utilizadas en muchos casos en la actualidad
  - *Gracias al Tte. Cnel. José Juan Lucena y a DAI-DGGC por los esfuerzos de captura de Ahumada III y por los scores para este ejemplo*

Caso simulado: rendimiento (precisión)

- Precisión del LR en condiciones de funcionamiento reales
  - En la línea de Daubert
- El perito utiliza BDRA y un protocolo adaptado al caso
- A incluir en el informe:
  - LR: peso de la evidencia
    - Probabilístico, basado en datos
    - Expresable en escalas verbales de apoyo
  - Precisión de los LR calculados en condiciones próximas a las del caso (por ejemplo, curva ECE)
    - Metodología científica
      - Repetible, empírica



Conclusiones

Conclusiones

- Reconocimiento automático de locutor
  - Estado del arte maduro con buen rendimiento y robustez
  - Pero aún quedan importantes retos que resolver…

Conclusiones

- Reconocimiento automático de locutor
  - Estado del arte maduro con buen rendimiento y robustez
  - Pero aún quedan importantes retos que resolver…
- Uso para valorar la evidencia forense: Metodología LR
    - “Emulando al ADN” [Gonzalez07]
    - Repetible, basada en datos
  - Arquitectura básica: transformación de score a LR

Conclusiones

- Reconocimiento automático de locutor
  - Estado del arte maduro con buen rendimiento y robustez
  - Pero aún quedan importantes retos que resolver…
- Uso para valorar la evidencia forense: Metodología LR
    - "Emulando al ADN" [Gonzalez07]
    - Repetible, basada en datos
  - Arquitectura básica: transformación de score a LR
- Importancia de la medida del rendimiento científico de las técnicas utilizadas
  - En condiciones de funcionamiento lo más próximas posibles a las reales
  - Necesidad de bases de datos de habla
    - BDRA, Baeza, Ahumada III, Ahumada IV… [Ramos08]

Conclusiones

- Reconocimiento automático de locutor
  - Estado del arte maduro con buen rendimiento y robustez
  - Pero aún quedan importantes retos que resolver…
- Uso para valorar la evidencia forense: Metodología LR
    - "Emulando al ADN" [Gonzalez07]
    - Repetible, basada en datos
  - Arquitectura básica: transformación de score a LR
- Importancia de la medida del rendimiento científico de las técnicas utilizadas
  - En condiciones de funcionamiento lo más próximas posibles a las reales
  - Necesidad de bases de datos de habla
    - BDRA, Baeza, Ahumada III, Ahumada IV… [Ramos08]
- Desafíos destacables
  - Selección de poblaciones (modelado de la hipótesis θ_d)
  - Adaptación a condiciones extremas en casos reales
  - Funcionamiento con pocos datos de habla

Referencias

Referencias

- [Reynolds00] D. A. Reynolds et al., 2000. “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, v. 10, pp. 19–41, 2000.
- [Campbell06] W. M. Campbell et al., 2006. “Support vector machines using GMM supervectors for speaker verification”. Signal Processing Letters, v. 13(5), pp. 308-311.
- [Reynolds03] D. A. Reynolds et al., 2003. “The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition”. Proc. of ICASSP 2003, Hong Kong, China.
- [Karajarekar04] S. Kajarekar et al., 2004. “Modelling NERFs for Speaker Recognition”. Proc. of Odyssey 2004, Toledo, Spain.
- [Stolcke06] A. Stolcke et al., 2005. “MLLR Transforms as Features in Speaker Recognition”. Proc. of Interspeech 2005, Lisbon, Portugal.
- [Rabiner07] L. Rabiner, 2007. “HMMs and Related Speech Technologies.” In Springer Handbook of Speech Technologies (ISBN: 978-3-540-49125-5). J. Benesty, M. M. Sondhi, Y. Huang (Eds.).

Referencias

- [Campbell06b] W. M. Campbell et al., 2006. “Support vector machines for speaker and language recognition”. Computer Speech and Language, v. 20(2-3), pp. 210-229.
- [Lopez07] I. Lopez-Moreno et al. “Support Vector Regression for Speaker Verification.” Proc. of Interspeech 2007, pp. 306-309. Antwerp, Belgium.
- [Auckenthaller00] R. Auckenthaler et al., 2000. “Score normalization for text-independent speaker verification systems.” Digital Signal Processing, vol. 10, pp. 42–54.
- [Brummer07] N. Brümmer et al., 2007. “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006.” IEEE Transactions on Audio, Speech and Signal Processing, vol. 15, no. 7, pp. 2072–2084.
- [Kenny07] P. Kenny et al., 2007. “Speaker and session variability in GMM-based speaker verification.” IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 4, pp. 1448–1460.

Referencias

- [Vogt07] R. Vogt and S. Sridharan, 2007. “Explicit modelling of session variability for speaker verification.” Computer Speech and Language, vol. 22, no. 1, pp. 17–38.
- [Vogt08] R. Vogt et al., 2008. “Factor Analysis Modelling for Speaker Verification with Short Utterances.” Proc. of Odyssey 2008, Stellenbosch, South Africa.
- [Fauve08] B. Fauve et al., 2008. “Improving the performance of text-independent short duration SVM- and GMM-based speaker verification.” Proc. Of Odyssey, Stellenbosch, South Africa.
- [Ramos08] D. Ramos et al., 2008. “Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish.” Proc. of Interspeech 2008, Brisbane, Australia.
- [Gonzalez07] J. Gonzalez-Rodriguez et al., 2007. “Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition,” IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no.7, pp. 2072–2084.