VI JORNADAS MAVIR

EXTRACCIÓN DE INFORMACIÓN DE AUDIO PARA RECUPERACIÓN DE INFORMACIÓN

Autor: Javier Franco Pedroso
javier.franco@uam.es

ATVS – Grupo de Reconocimiento Biométrico
http://atvs.ii.uam.es

UNIVERSIDAD AUTONOMA DE MADRID

ATVS

Índice

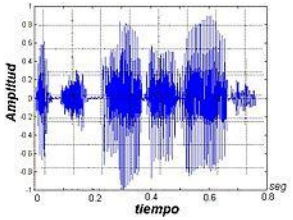
Índice

1.1. Información en la señal de audio

- **Voz:**
  - Identidad del hablante.
  - Estado emocional del hablante.
  - Género del hablante.
  - Mensaje.
  - Idioma hablado.
- **Música:**
  - Estilo, ritmo, melodía, etc.
  - Pieza concreta.
- **Otros:**
  - Detección de eventos sonoros concretos

1.1. Información en la señal de audio

- **Voz:**
  - Identidad del hablante.
  - Estado emocional del hablante.
  - Género del hablante.
  - Mensaje.
  - Idioma hablado.



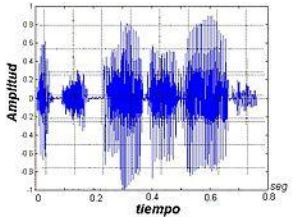
- **Música:**
  - Estilo, ritmo, melodía, etc.
  - Pieza concreta.

- **Otros:**
  - Detección de eventos sonoros concretos

1.1. Información en la señal de audio

- **Voz:**
  - Identidad del hablante.
  - Estado emocional del hablante.
  - Género del hablante.
  - Mensaje.
  - Idioma hablado.
- **Música:**
  - Estilo, ritmo, melodía, etc.
  - Pieza concreta.
- **Otros:**
  - Detección de eventos sonoros concretos

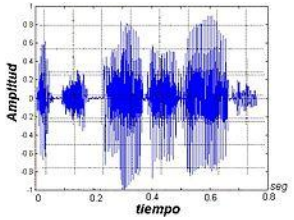
Segmentación de audio



1.1. Información en la señal de audio

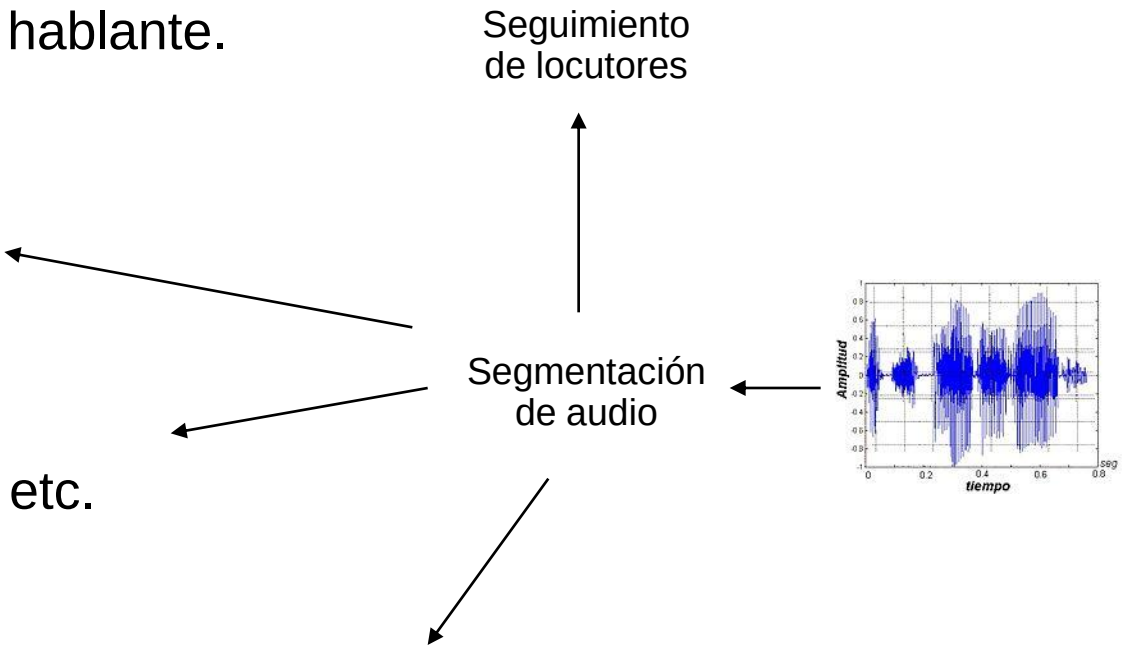
- **Voz:**
  - Identidad del hablante.
  - Estado emocional del hablante.
  - Género del hablante.
  - Mensaje.
  - Idioma hablado.
- **Música:**
  - Estilo, ritmo, melodía, etc.
  - Pieza concreta.
- **Otros:**
  - Detección de eventos sonoros concretos

Segmentación de audio



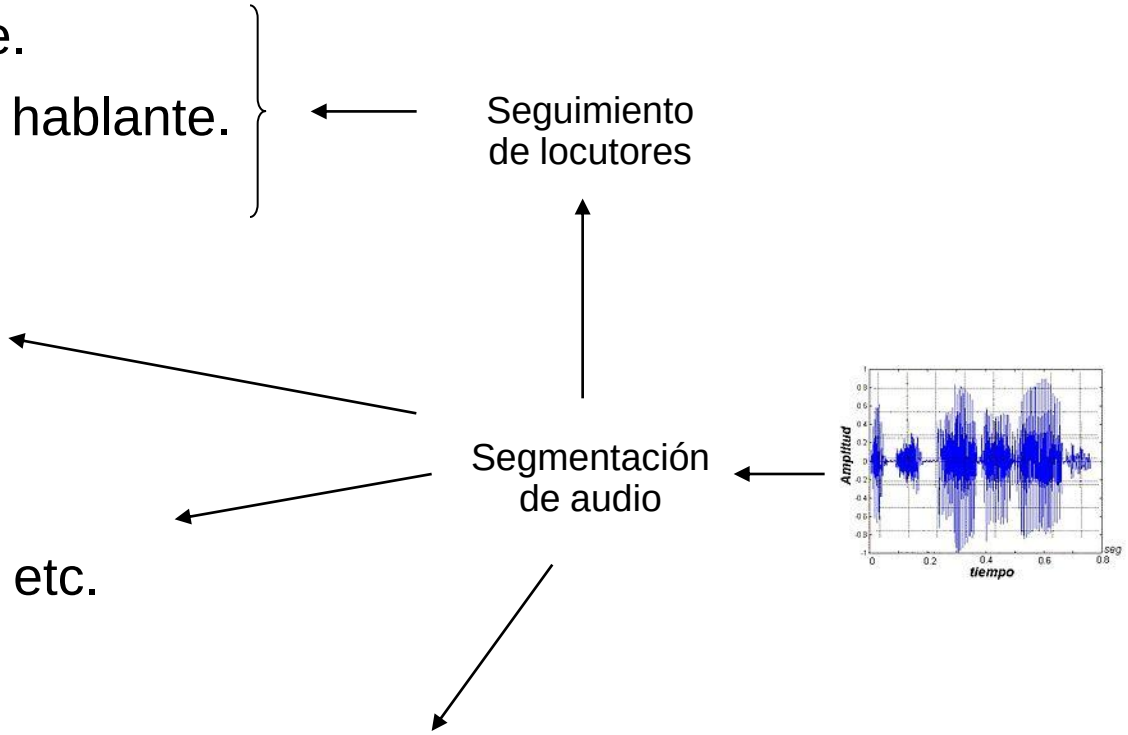
1.1. Información en la señal de audio

- **Voz:**
  - Identidad del hablante.
  - Estado emocional del hablante.
  - Género del hablante.
  - Mensaje.
  - Idioma hablado.
- **Música:**
  - Estilo, ritmo, melodía, etc.
  - Pieza concreta.
- **Otros:**
  - Detección de eventos sonoros concretos



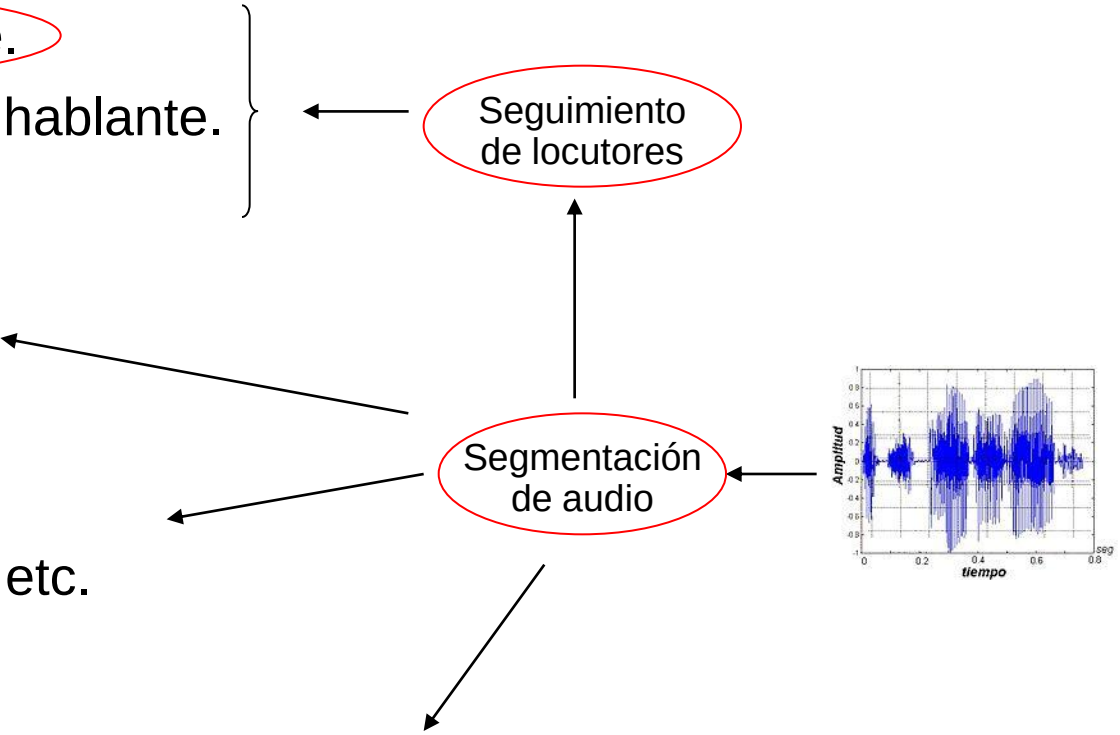
1.1. Información en la señal de audio

- **Voz:**
  - Identidad del hablante.
  - Estado emocional del hablante.
  - Género del hablante.
  - Mensaje.
  - Idioma hablado.
- **Música:**
  - Estilo, ritmo, melodía, etc.
  - Pieza concreta.
- **Otros:**
  - Detección de eventos sonoros concretos



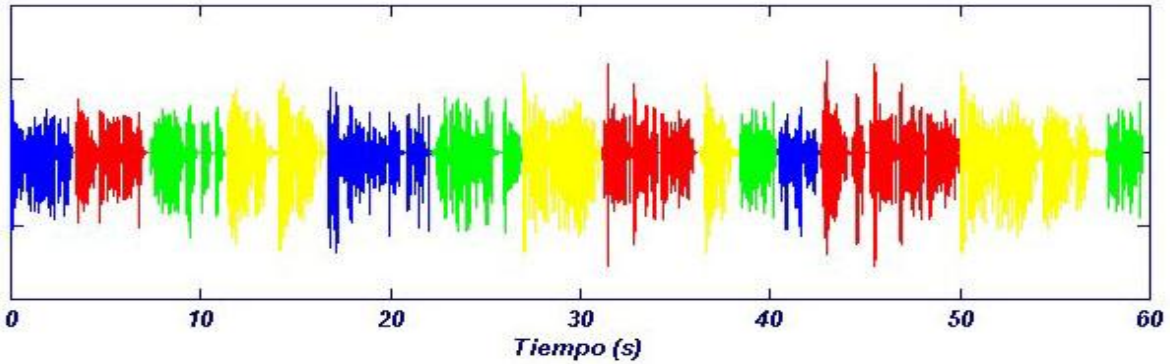
1.1. Información en la señal de audio

- **Voz:**
  - Identidad del hablante.
  - Estado emocional del hablante.
  - Género del hablante.
  - Mensaje.
  - Idioma hablado.
- **Música:**
  - Estilo, ritmo, melodía, etc.
  - Pieza concreta.
- **Otros:**
  - Detección de eventos sonoros concretos



1.2. Segmentación de audio

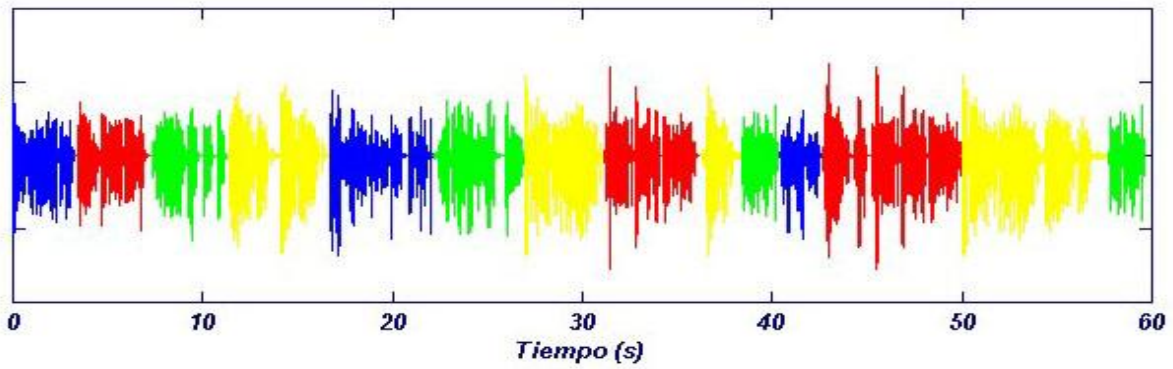
- **Definición:** determinar los límites de segmentos de audio homogéneos en base a categorías genéricas.
  - Voz
  - Música
  - Voz + ruido
  - Otros



- **Objetivos:**
  - Primer nivel de indexación (*browsing*).
  - Post-procesado específico de cada categoría de audio.
    - Voz: seguimiento de locutores, transcripción automática, *word-spotting*, …
    - Música: *Music Information Retrieval* (MIR)
    - Otros: identificación de eventos sonoros.

1.3. Seguimiento de locutores

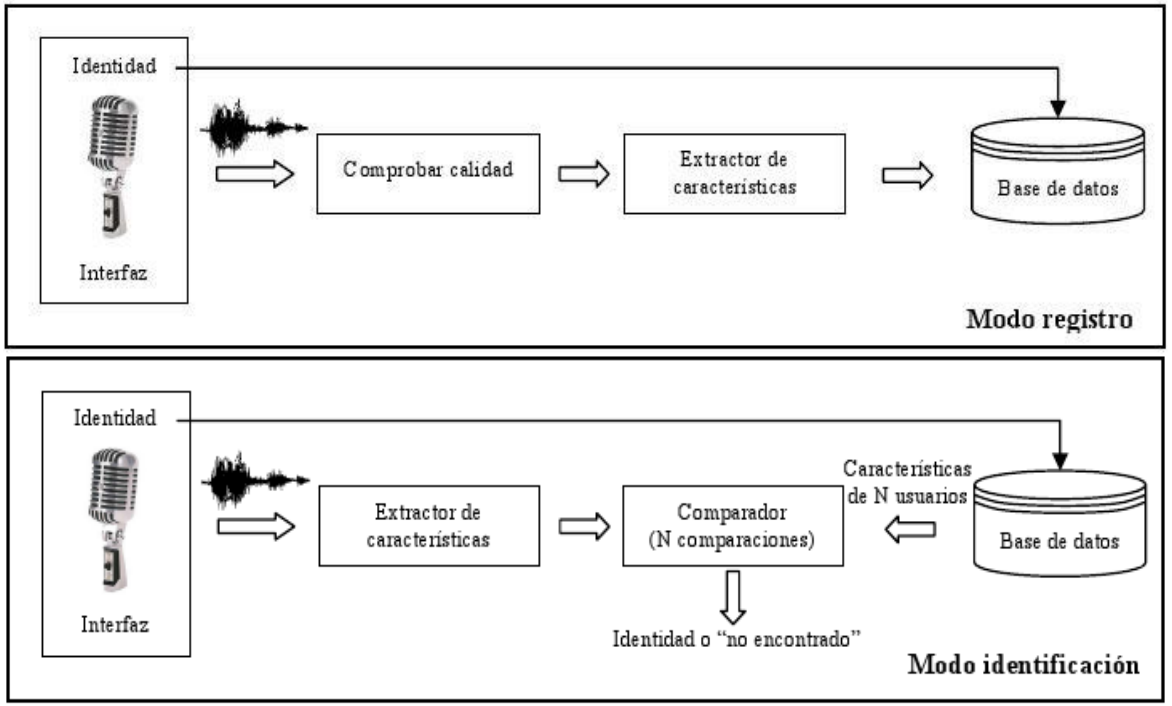
- **Definición:** determinar los intervalos de tiempo en que intervienen distintas personas en un segmento de voz dado.
  - Locutor A
  - Locutor B
  - Locutor C
  - Locutor D



- **Objetivos:**
  - Navegación entre locutores.
  - Permitir la identificación de cada locutor.
  - Permitir adaptación al locutor para el reconocimiento de habla.

1.4. Identificación de locutor

- **Definición:** determinar la identidad del hablante en un segmento de voz dado.



- **Objetivos:**
  - Indexación de contenidos por locutor.
  - Disponer de más datos para adaptación al locutor.

1.5. Identificación de idioma

- **Definición:** determinar el idioma hablado en un segmento de voz dado.
- **Objetivos:**
  - Indexación de contenidos por idioma.
  - Aplicar el reconocedor de habla del idioma apropiado.

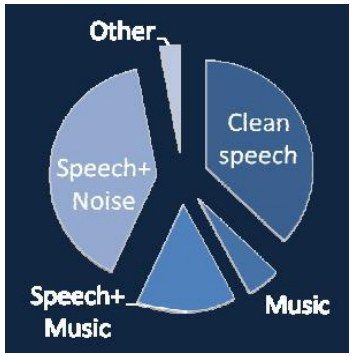
Índice

2. Evaluaciones Albayzín

- Organizadas por la Red Temática en Tecnologías del Habla (RTTH) española.
  - Foro de investigación financiado por el MEC.
- Bianuales desde 2006. Cada edición engloba un conjunto de disciplinas diferente:
  - 2006: mimetización de voces, segmentación e identificación de locutores, traducción texto a lengua de signos.
  - 2008: verificación de la lengua, síntesis de voz, traducción Castellano-Euskera.
  - 2010: identificación de idioma, síntesis de voz, segmentación de audio, seguimiento de locutores.
- Participación de grupos de investigación a nivel nacional principalmente.

2.1. Evaluación de segmentación de audio 2010 (1)

- Tarea: segmentar en 5 clases acústicas
  - Voz limpia
  - Voz + ruido
  - Voz + música
  - Música
  - Otros



Martin Zelenák, Henrik Schulz and Javier Hernando. "Albayzin 2010 Evaluation Campaign: Speaker Diarization"

- Base de datos: programas de noticias en catalán (TV3/24).
  - Desarrollo: 16 grabaciones de ~4 horas cada una.
  - Evaluación: 8 grabaciones de ~4 horas cada una.
- Medida de rendimiento: error de clasificación promedio

$$Error = \underset{i}{average}\left(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}\right)$$

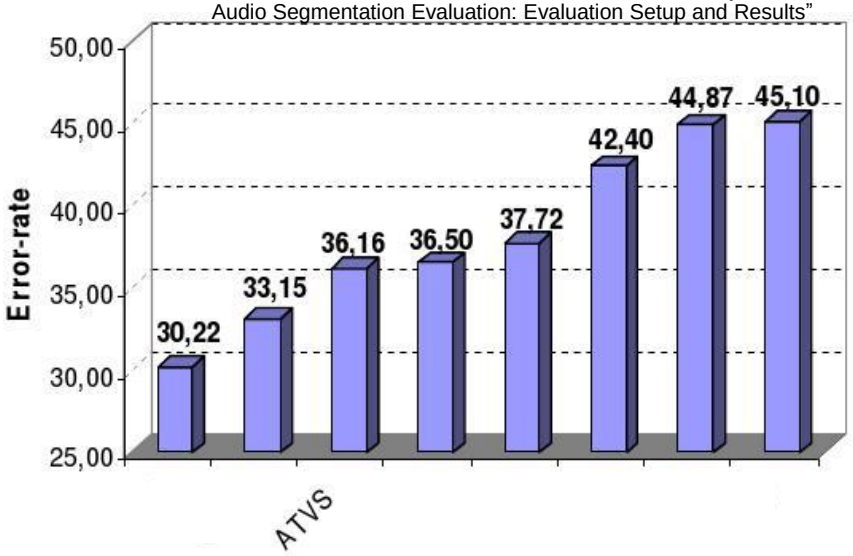
Taras Butko, Climent Nadeu and Henrik Schulz. "Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results"

2.1. Evaluación de segmentación de audio 2010 (2)

- Sistema ATVS:
  - Decodificación de Viterbi mediante HMM de 5 estados.
  - Un GMM de 1024 gaussianas por estado.
  - Entrenamiento discriminativo de los GMM's.
- Resultados por clase y ranking de la evaluación:

Data set	Misclassification error-rate (%)				
	'mu'	'sp'	'sm'	'sn'	Final
Development	18.43	22.69	16.86	25.30	20.89
Evaluation	31.01	40.42	33.39	39.80	36.16

Taras Butko, Climent Nadeu and Henrik Schulz. "Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results"



2.2. Evaluación de seguimiento de locutor 2010 (1)

- Tarea: determinar intervalos de tiempo en que hablan distintos locutores.
- Base de datos: misma que para segmentación de audio.
  - 66-120 locutores por grabación.
- Medida de rendimiento: *diarization error rate* (DER)
  - Alineación uno a uno entre salida del sistema y referencia.

Martin Zelenák, Henrik Schulz and Javier Hernando. “Albayzin 2010 Evaluation Campaign: Speaker Diariization”

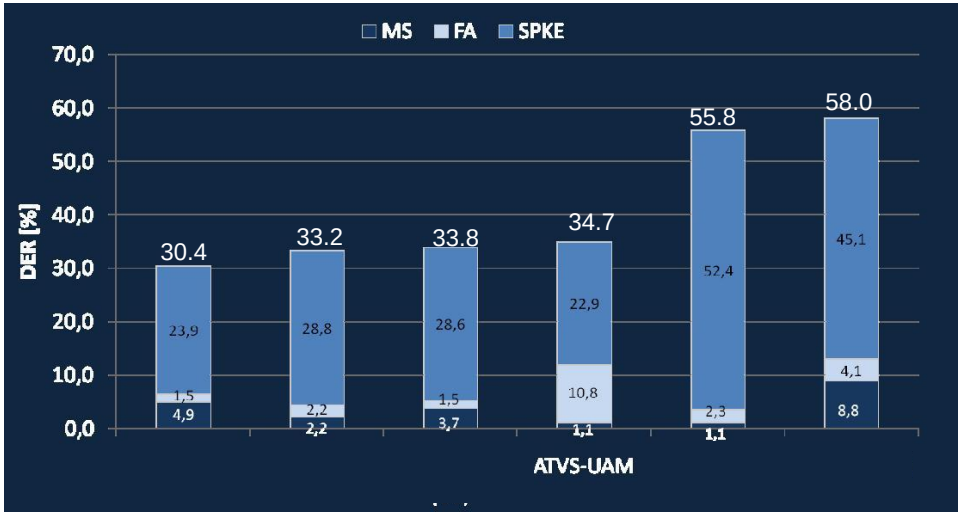
  - (duración errores “diarizacion”) / (duración segmento evaluado)

2.2. Evaluación de seguimiento de locutor 2010 (2)

- Sistema ATVS:
  - *Front-end* de *factor analysis* para extracción de *i-vectors*.
  - Agrupamiento de *i-vectors* mediante distancia coseno.
  - Alineamiento de locutores mediante decodificación de Viterbi.
- Resultados y ranking de la evaluación:

Martin Zelenák, Henrik Schulz and Javier Hernando. "Albayzin 2010 Evaluation Campaign: Speaker Diarization"

Data set	Missed (%)	False Alarm (%)	Speaker error (%)	DER (%)
Develop.	7.8	10.3	16.5	34.5
Eval.	1.1	10.8	22.9	34.7



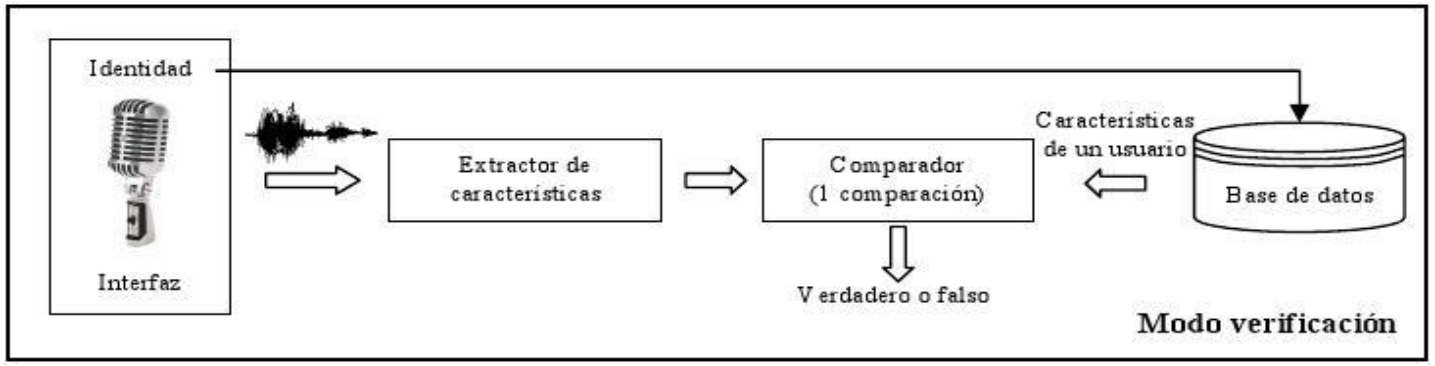
Índice

3. Evaluaciones NIST

- Organizadas por *National Intitute of Standards and Technology* (NIST) estadounidense.
  - Promover la innovación y la competencia industrial en EEUU.
- *Multimodal Information Group*: enfocado a la información multimedia y multilingüe.
  - *Speaker Recognition Evaluation* (SRE).
  - *Language Recognition Evaluation* (LRE).
  - *Rich Transcription Evaluation* (ASR, STT, *Diarization*, …).
  - TRECVid *Surveillance/Multimedia Event Detection Evaluation*.
- NIST SRE: anuales entre 1996 y 2006. Bianuales 2006-…
- NIST LRE: bianuales desde 2003 (primera en 1996).

3.1. NIST SRE 2010 (1)

- Tarea: verificación de locutor.
  - Distintas condiciones de entrenamiento (10 s a 8 conversaciones 2.5') y test (10 s a 1 conversación).

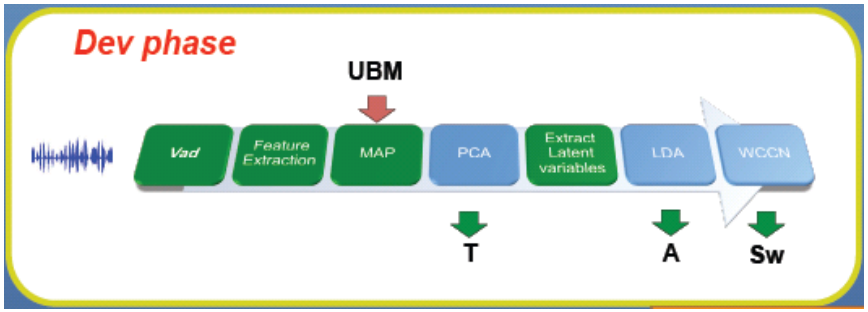


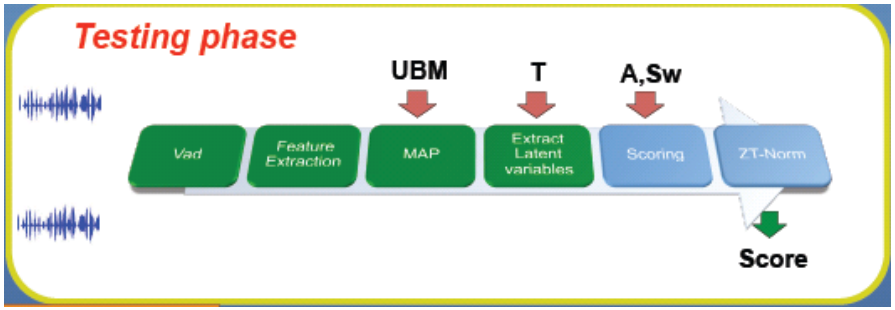
- Base de datos:
  - Tipos de grabaciones: distintos canales telefónicos y microfónicos.
  - Centenas de locutores, cientos de miles de comparaciones.
- Medida de rendimiento: *Detection Cost Function* (DCF)

$$C_{Det} = C_{Miss} \cdot P_{Miss|Target} \cdot P_{Target} + C_{FalseAlarm} \cdot P_{FalseAlarm|NonTarget} \cdot (1 - P_{Target})$$

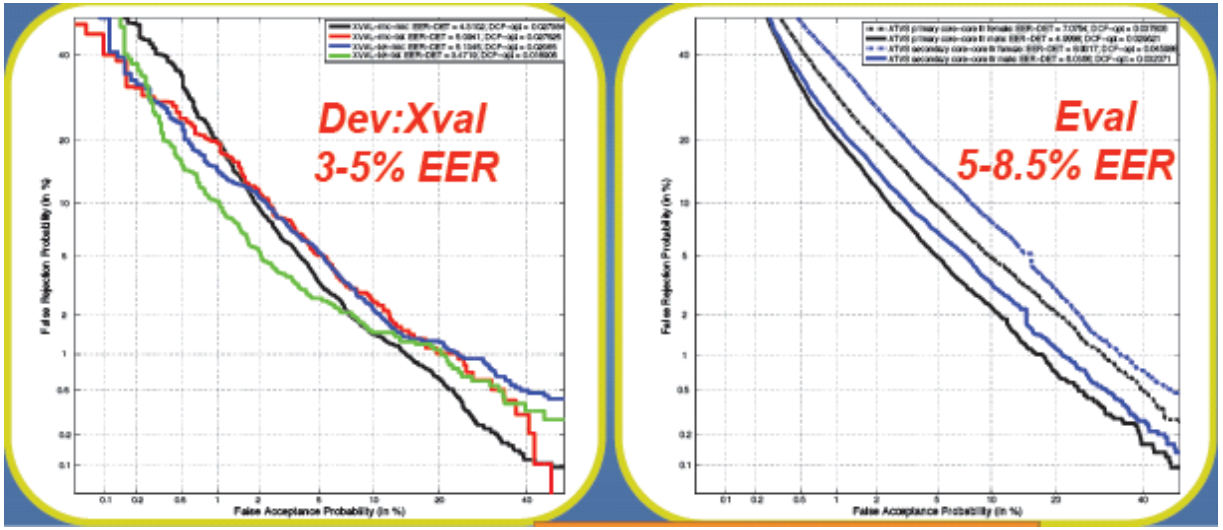
3.1. NIST SRE 2010 (2)

- Sistema ATVS: *total variability*





- Resultados:

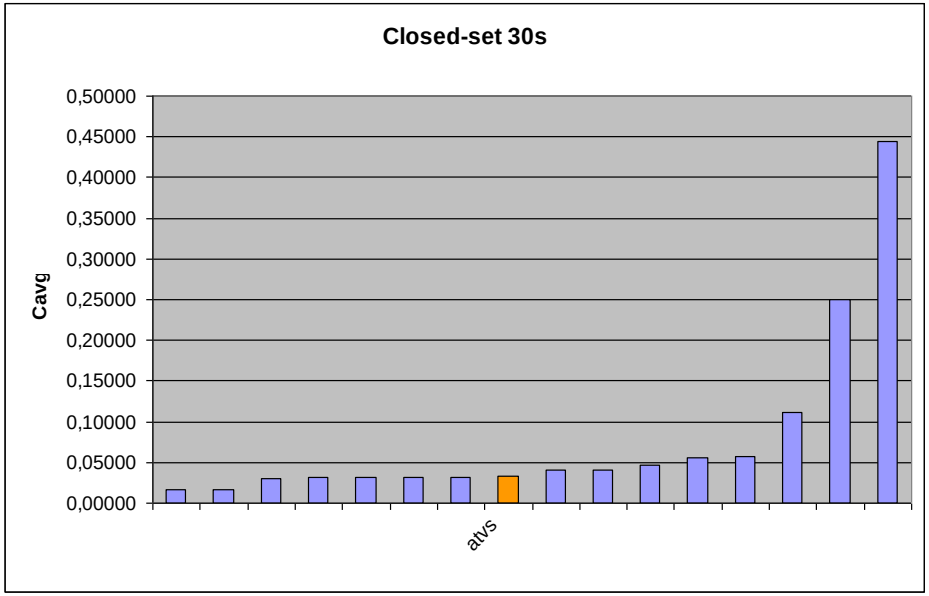


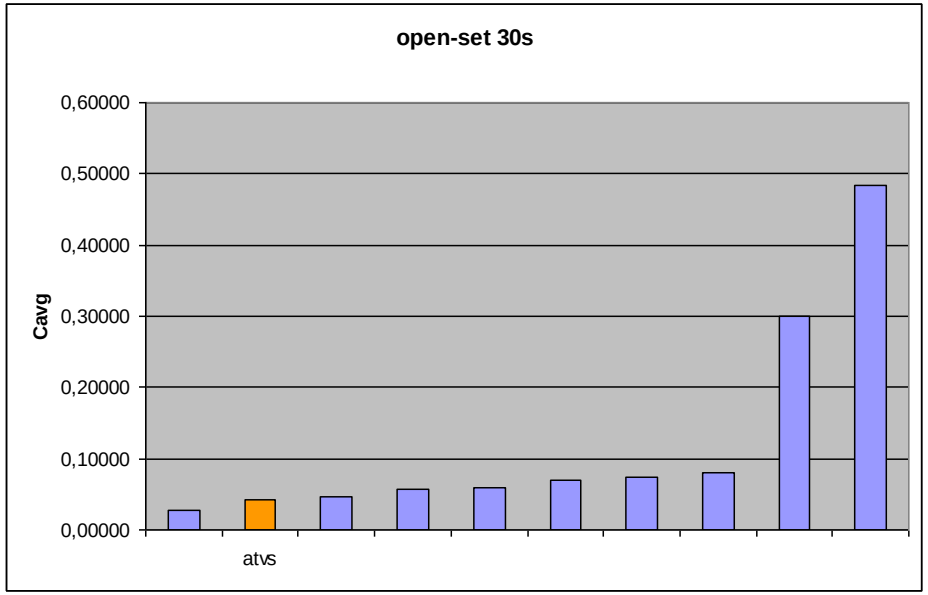
3.2. NIST LRE 2009 (1)

- Tarea: verificación de idioma.
  - Datos de entrenamiento “ilimitados”.
  - Distintas longitudes de test: 3, 10 y 30 segundos.
  - Evaluación en conjunto abierto y cerrado.
- Base de datos
  - Dos tipos de grabaciones: conversaciones telefónicas y llamadas a programas radiofónicos (Voice Of America).
  - 23 idiomas *target*
- Medida de rendimiento: coste promediado sobre el conjunto de idiomas *target* (C_{avg}).
  - Distinto para conjunto abierto y cerrado.

3.2. NIST LRE 2009 (2)

- Sistema ATVS: fusión de
  - Sistemas acústicos: FA-GMM y SVM-SV
  - Sistemas fonotácticos: PhoneSVM
  - Sistemas TNO: FA-GMM y SVM-SV
- Resultados





Muchas gracias