rtve

Instituto Universitario de Investigación en Ingeniería de Aragón
Universidad Zaragoza

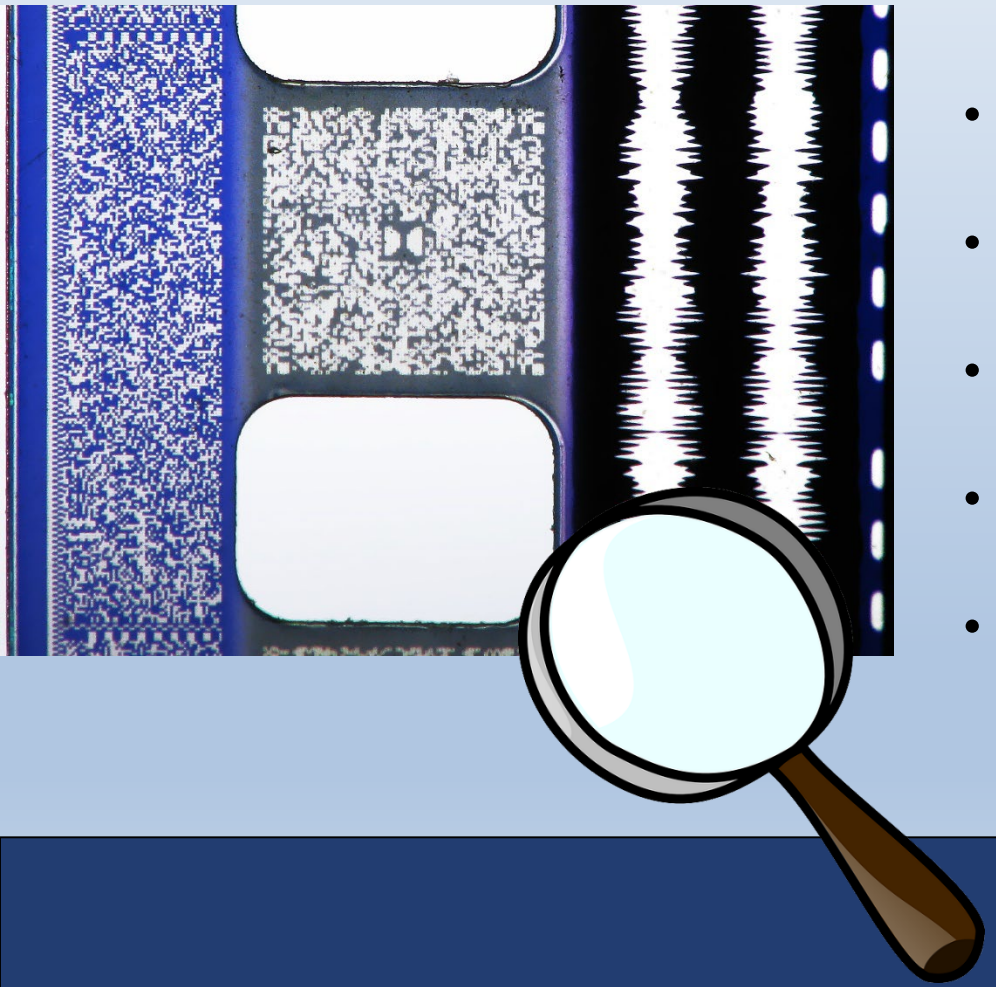
Tecnologías para el análisis y extracción de metadatos en contenidos audiovisuales: Tecnologías del Habla

VIVOLAB

Universidad Zaragoza
1542

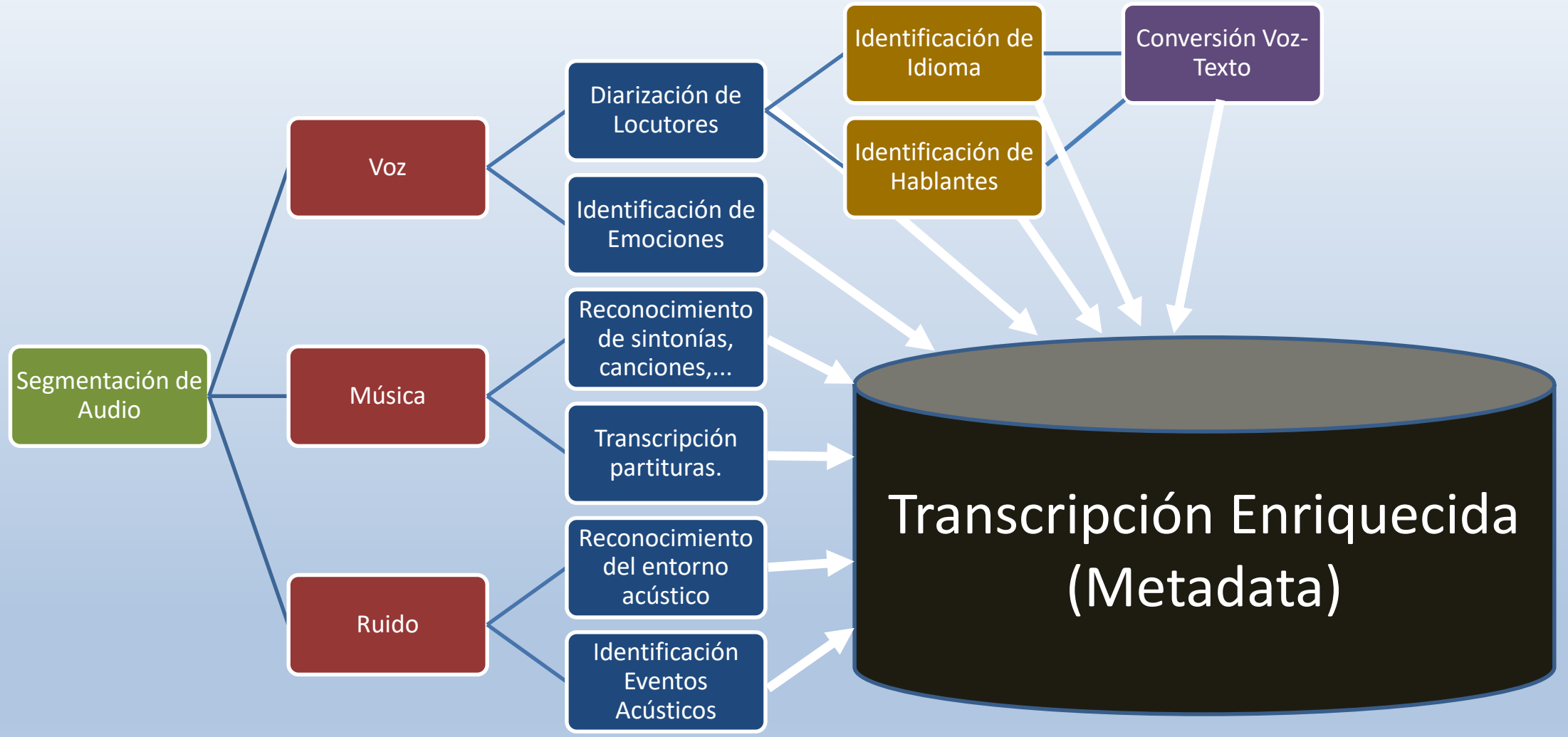
Extracción de Información: Audio

¿Qué información podemos encontrar en un audio?

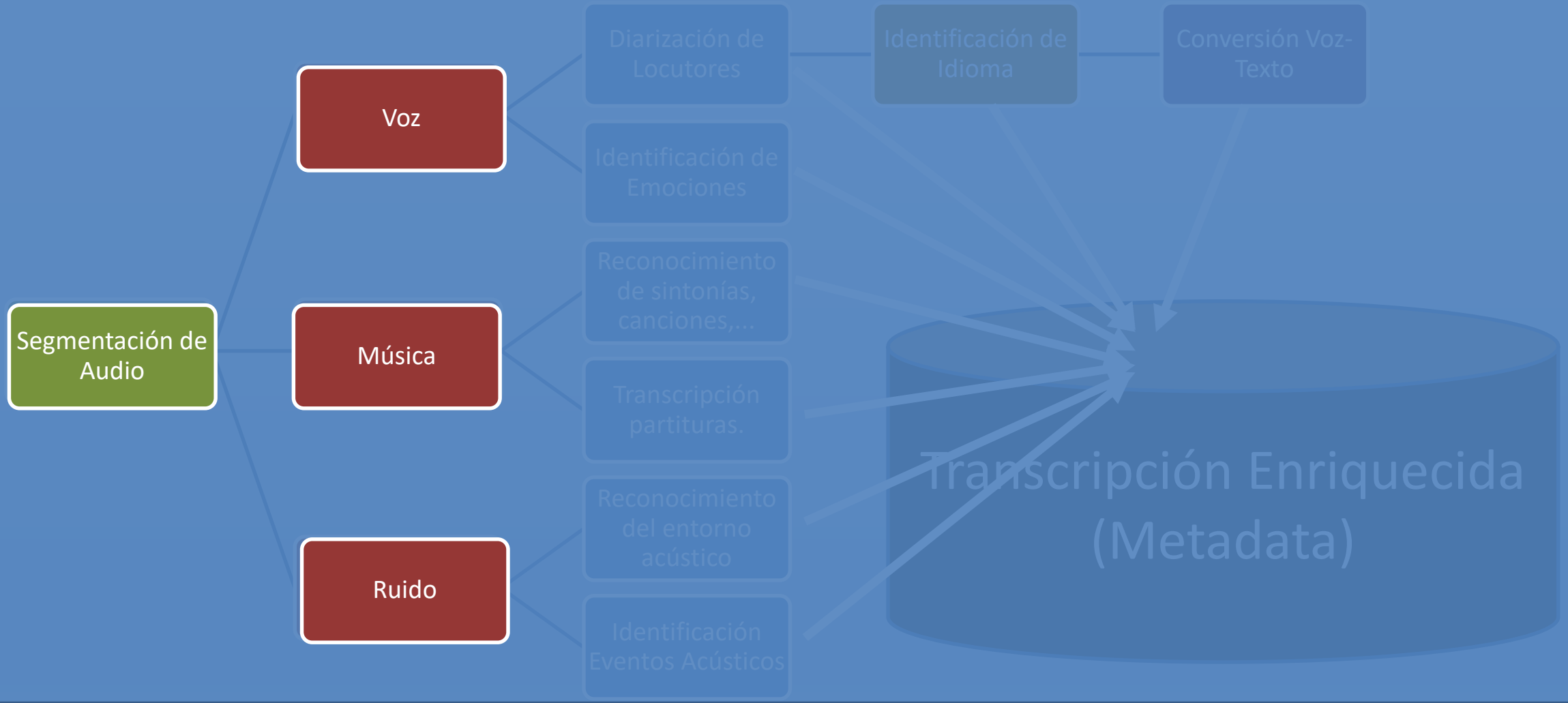


- Hay ruido, música, habla, …
- Cuántas personas hablan y cuándo habla cada una de ellas
- Cuáles son las identidades de las personas que hablan
- En qué idioma están hablando
- Qué dice cada una de ellas
- Cuál es el estado emocional de cada una de ellas

Tecnologías



Segmentación de Audio



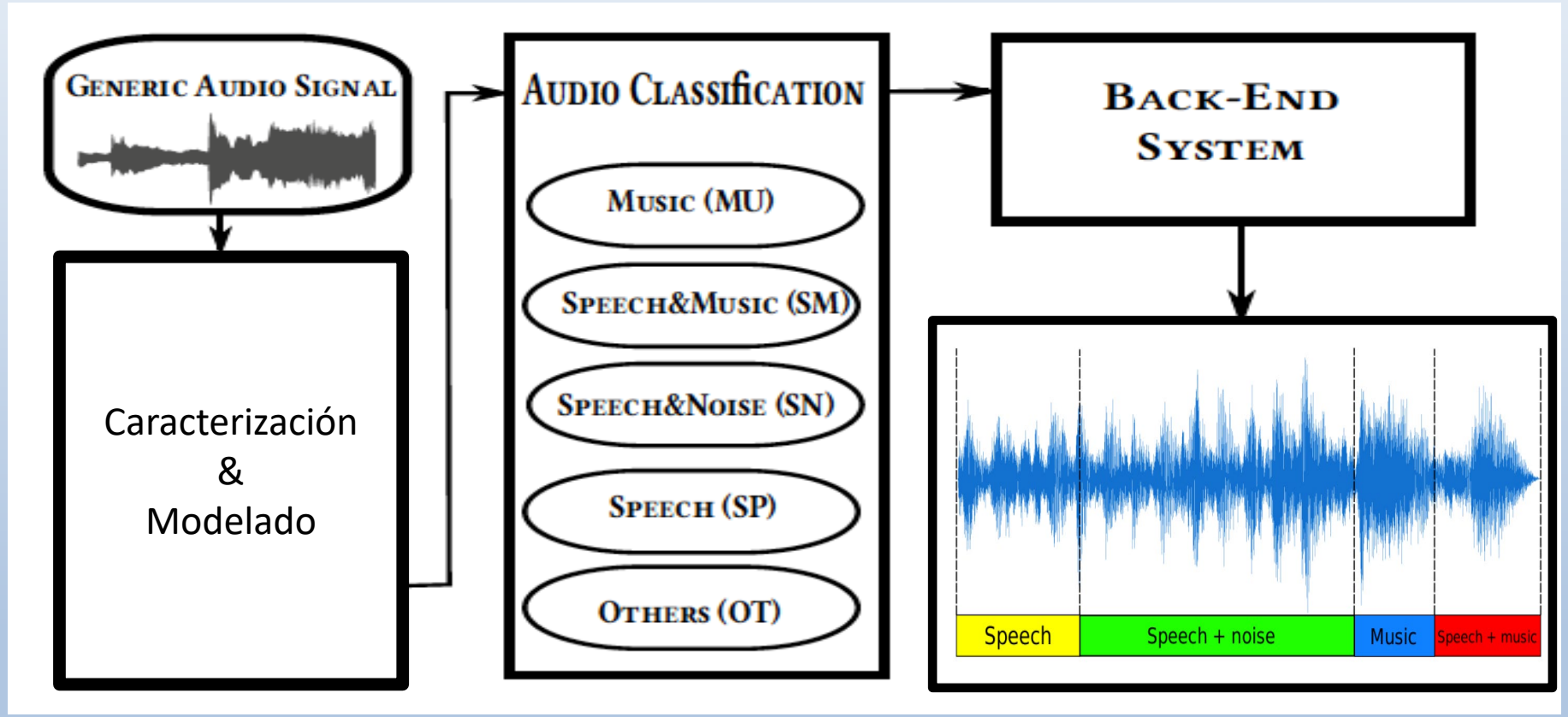
Segmentación de Audio

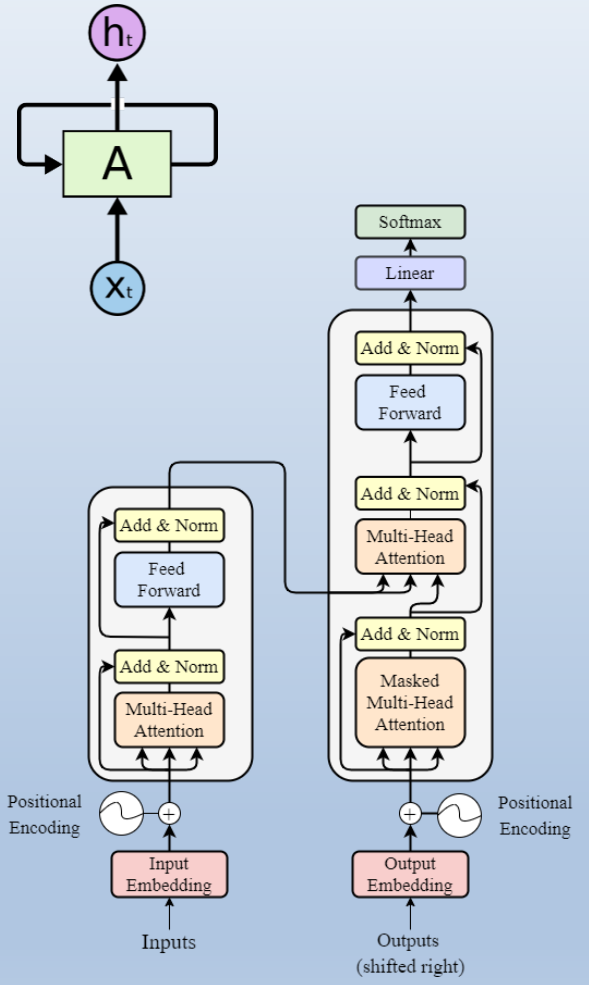
- **¿Qué es?:**
  - Dividir el audio de entrada en fragmentos atendiendo al tipo de contenido acústico: Voz / Música / Ruido y combinaciones de estos.



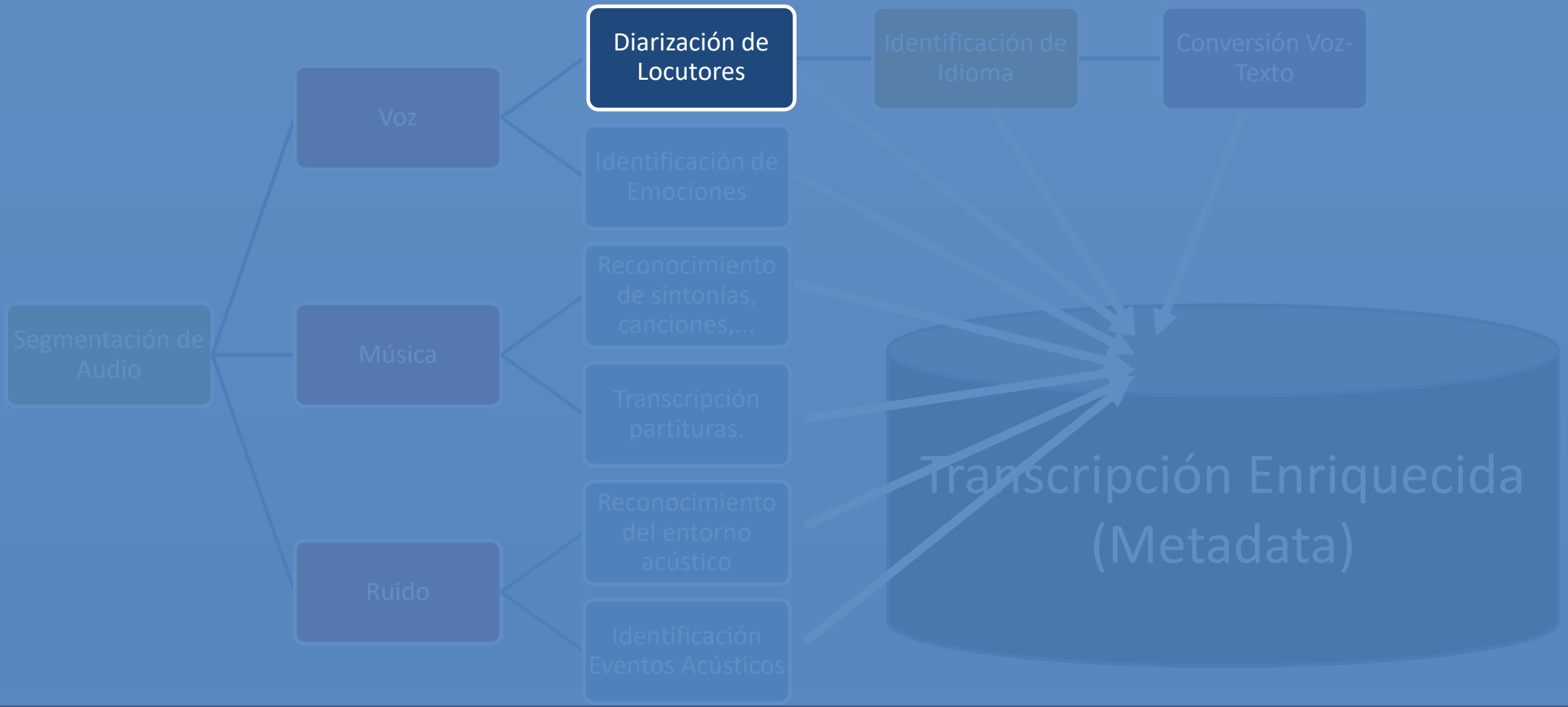
- **¿Para qué sirve?:**
  - Da soporte a otras tareas de extracción de información como:
    - Diarización
    - Identificación del hablante
    - Conversión Voz-Texto
    - …

Segmentación de Audio



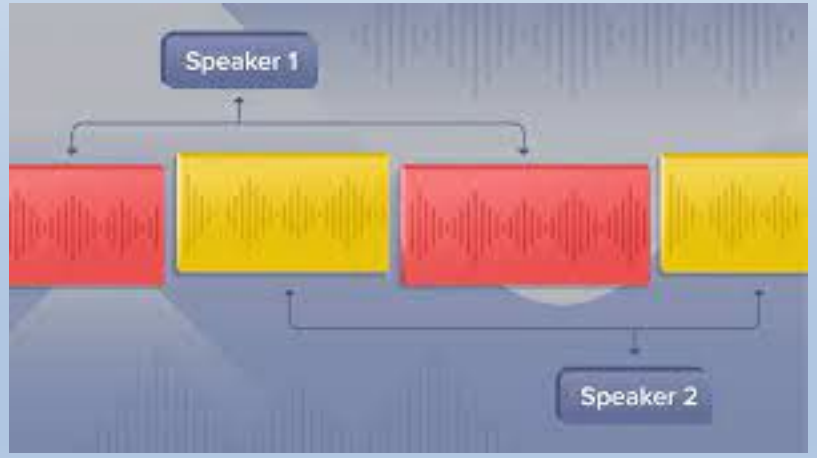


Segmentación y Agrupación de Hablantes



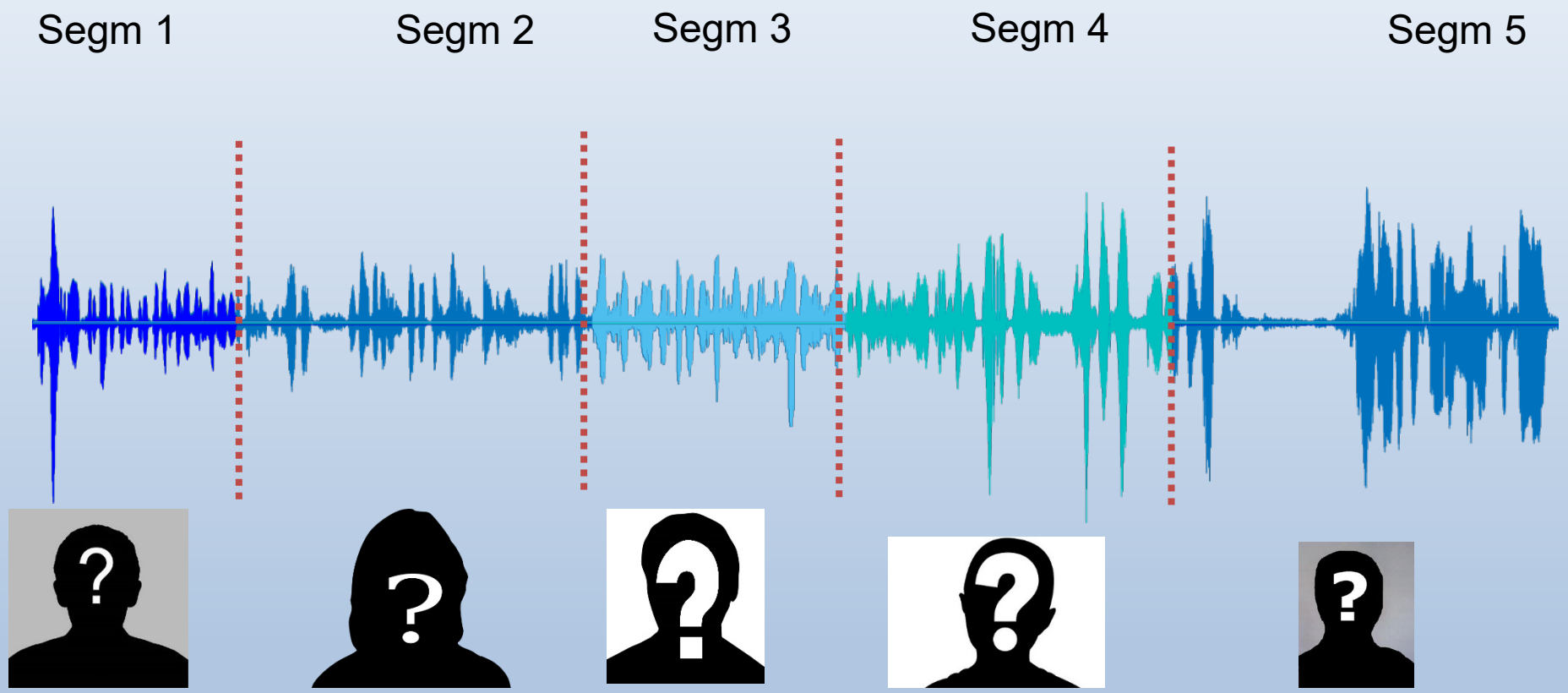
Segmentación y Agrupación de Hablantes

- **¿Qué es?:**
  - **Dividir en fragmentos atendiendo al interviniente y agrupar dichos fragmentos en función de la identidad del locutor.**
  - **Término usado por la comunidad: Diarización**
    - *Diarise: (Diarize) to make use of a diary to record past events or those planned for the future.*
- **¿Para qué sirve?:**
  - **Tecnología soporte para mejorar prestaciones**
    - **Reconocimiento automático del habla**
    - **Reconocimiento del hablante**
    - **…**

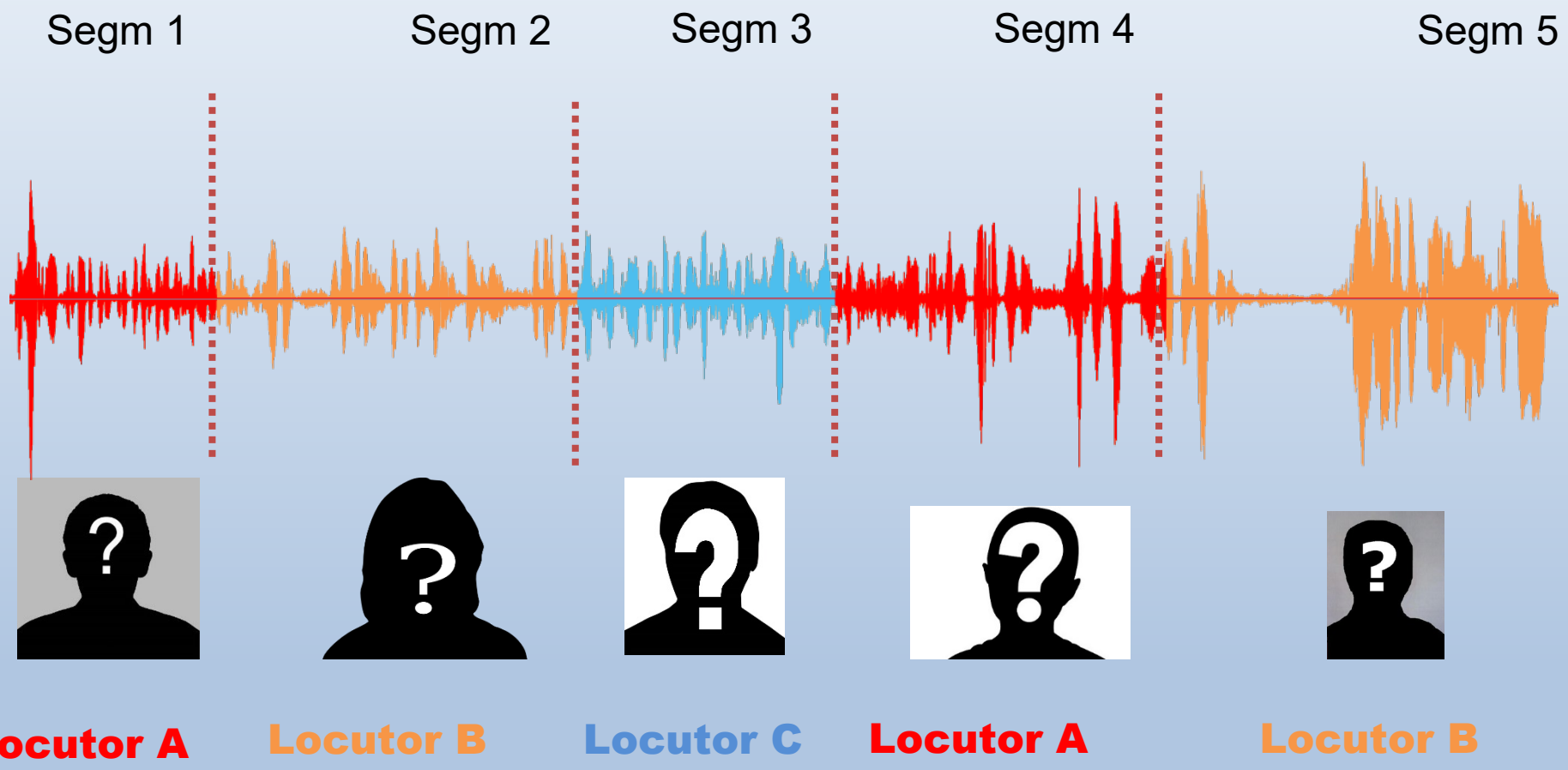


Segmentación y Agrupación de Hablantes

Segmentación y Agrupación de Hablantes

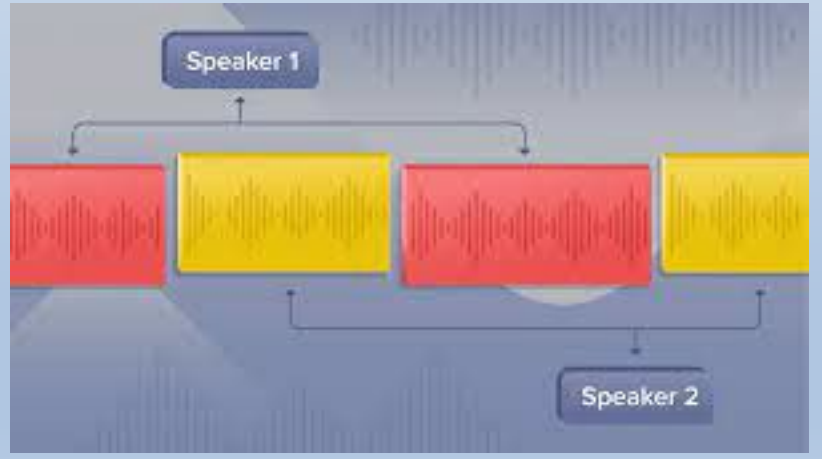


Segmentación y Agrupación de Hablantes



Segmentación y Agrupación de Hablantes

- **¿Para qué sirve?:**
  - Tecnología soporte para mejorar prestaciones de:
    - Reconocimiento automático del habla
    - Reconocimiento del hablante
    - …

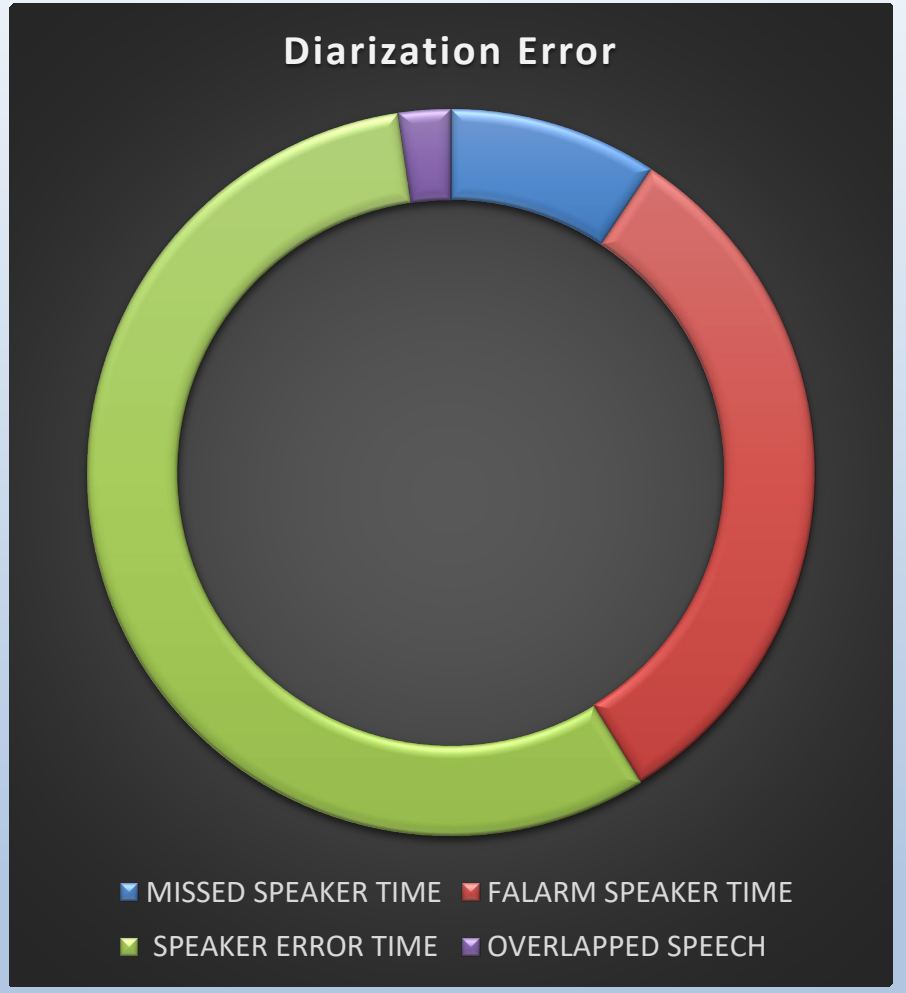


Medida del Error

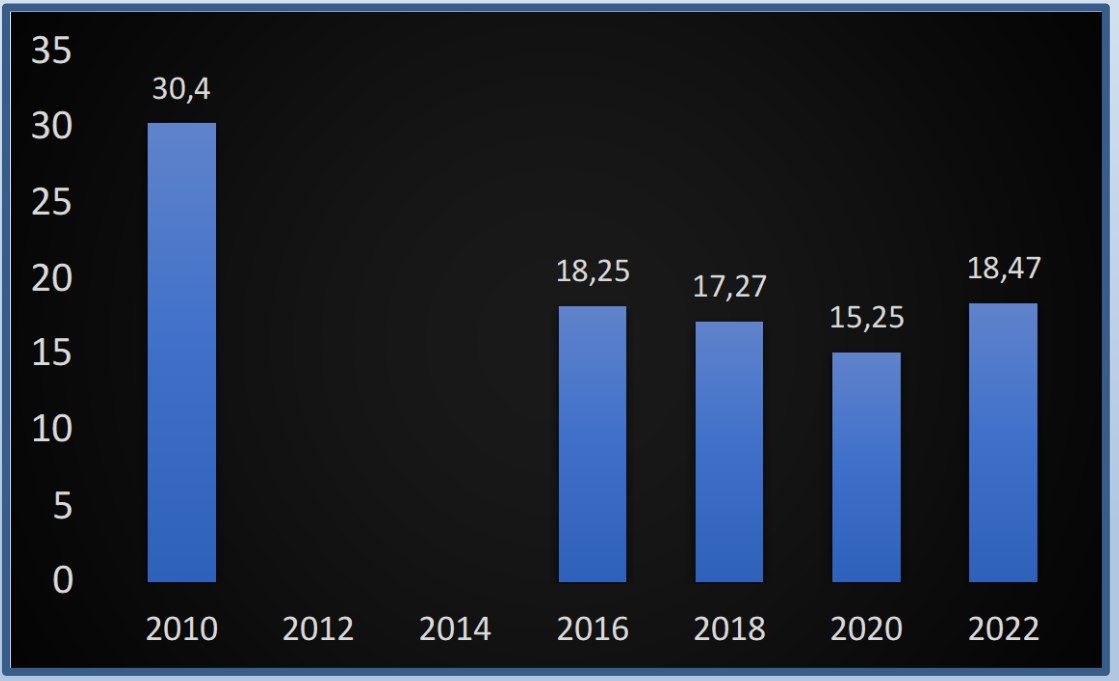
- **Diarization Error Rate:**

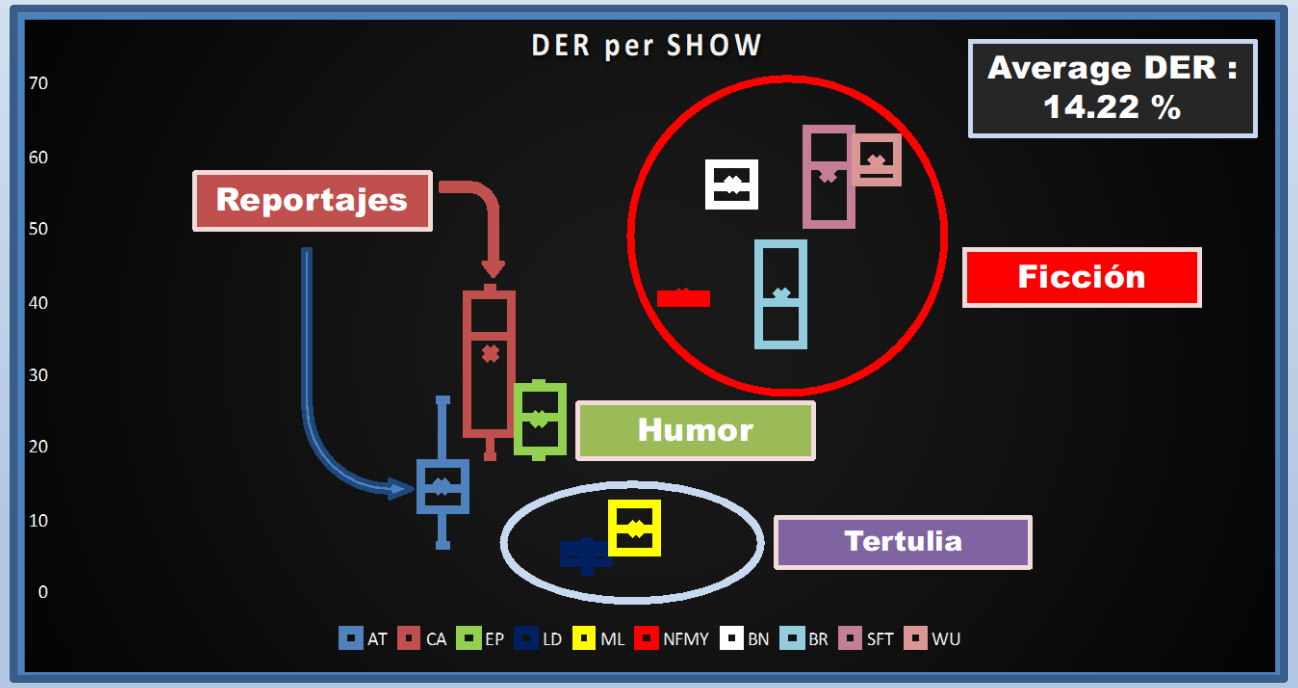
$$DER = \frac{T_{incorrecto}}{T_{voz}}$$

  - Componentes del error:
    - Pérdida:
      - Hay voz pero se ha confundido con silencio.
    - Falsa Alarma:
      - No hay voz pero se ha detectado erróneamente.
    - Error de Locutor:
      - Se ha confundido un locutor con otro.
    - Error por Solape:
      - Dos locutores hablan a la vez, pero solo se ha identificado a uno.



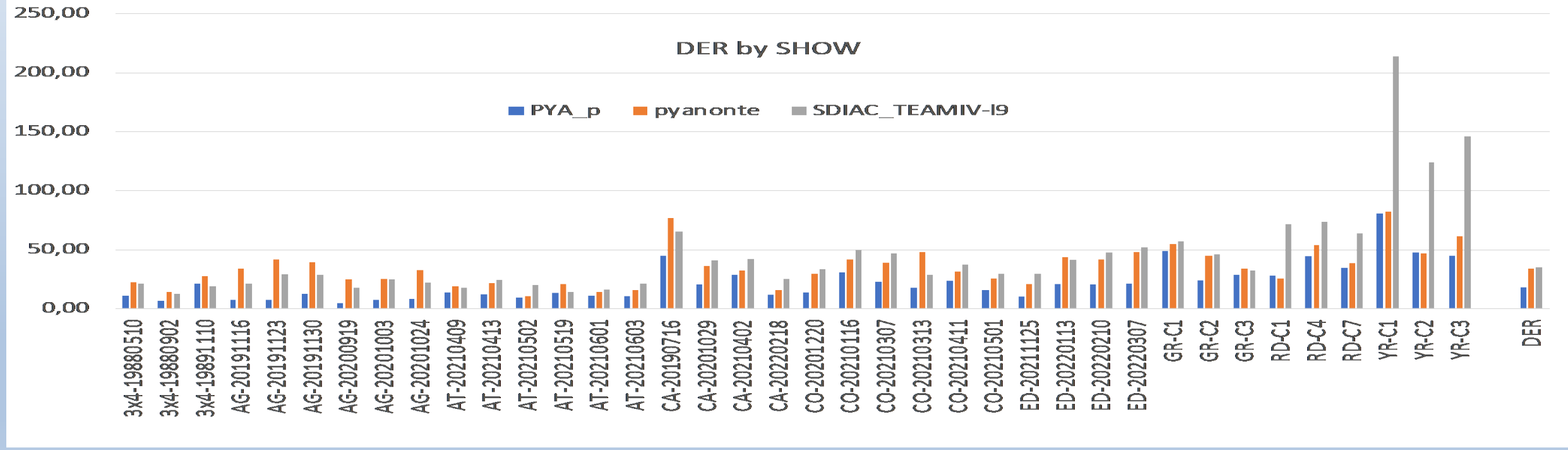
Prestaciones: Albayzín Retos - RTVE



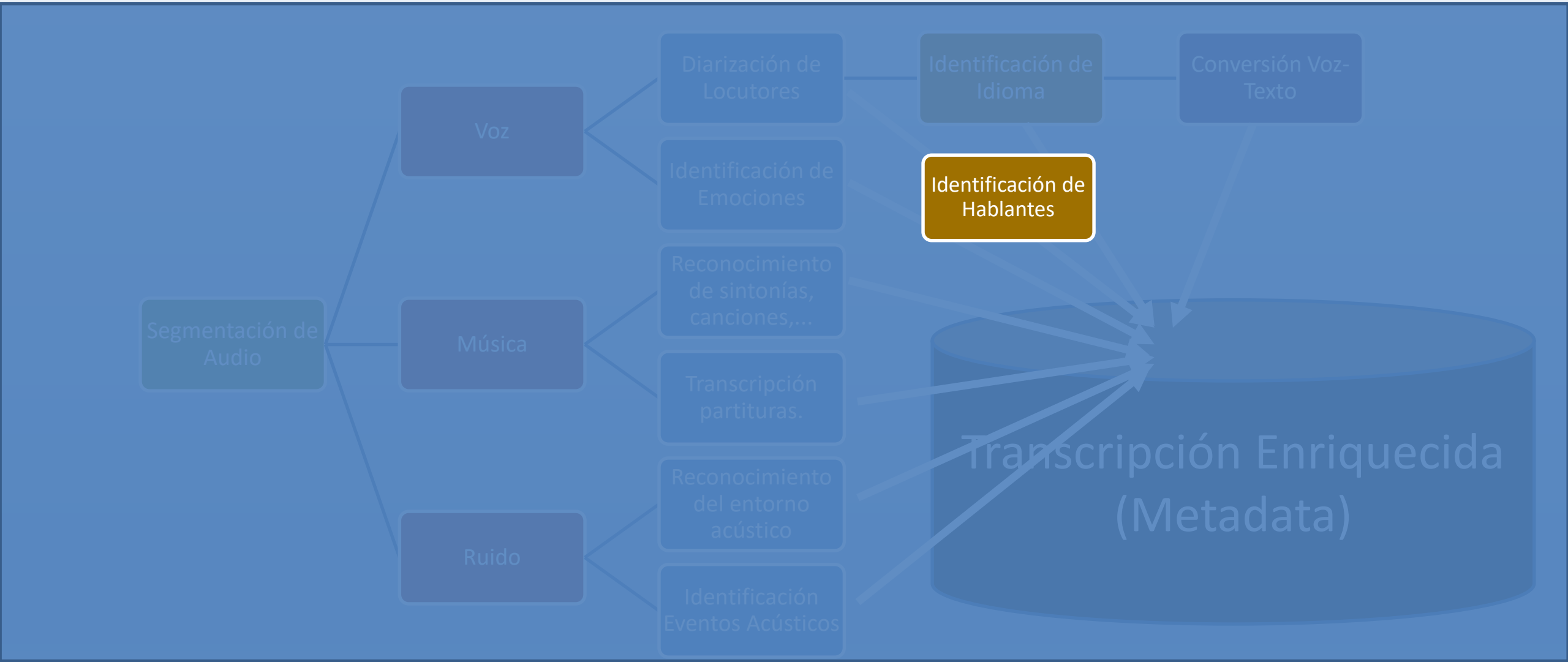


Prestaciones: Albayzín Retos - RTVE

https://github.com/pyannote/pyannote-audio



Identificación de Hablantes

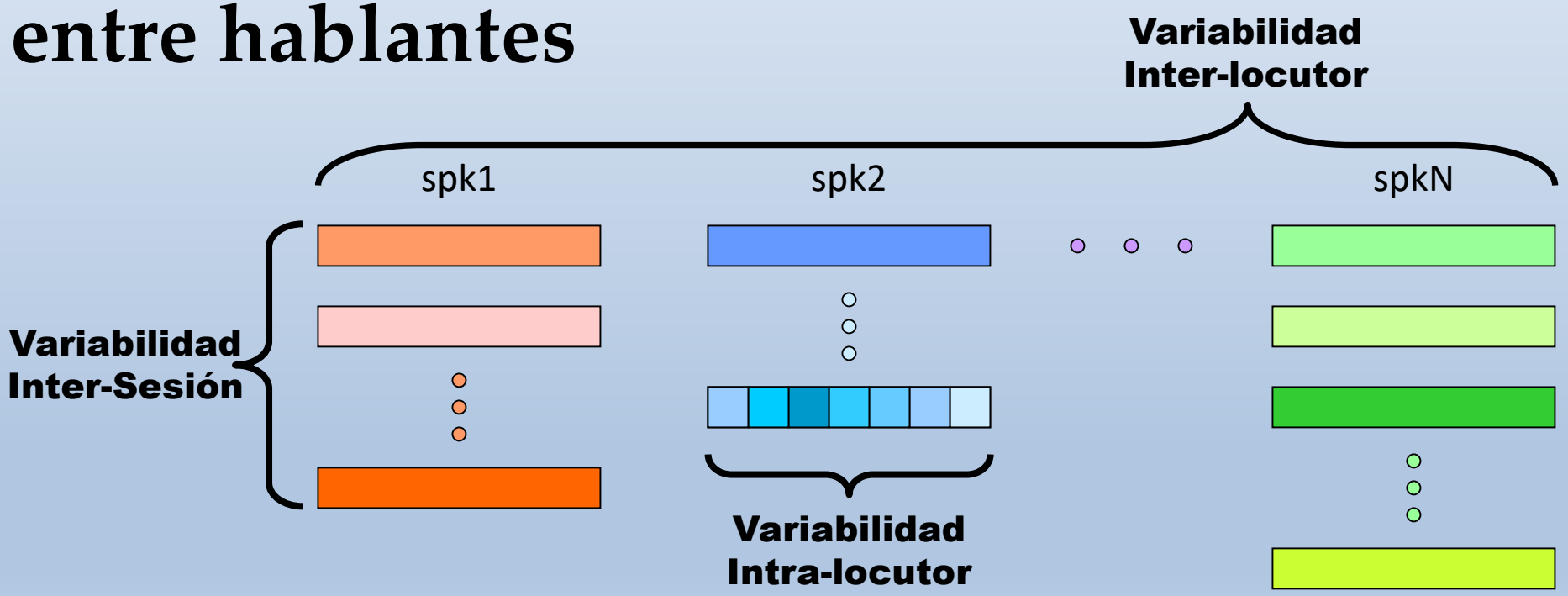


Identificación de Hablantes

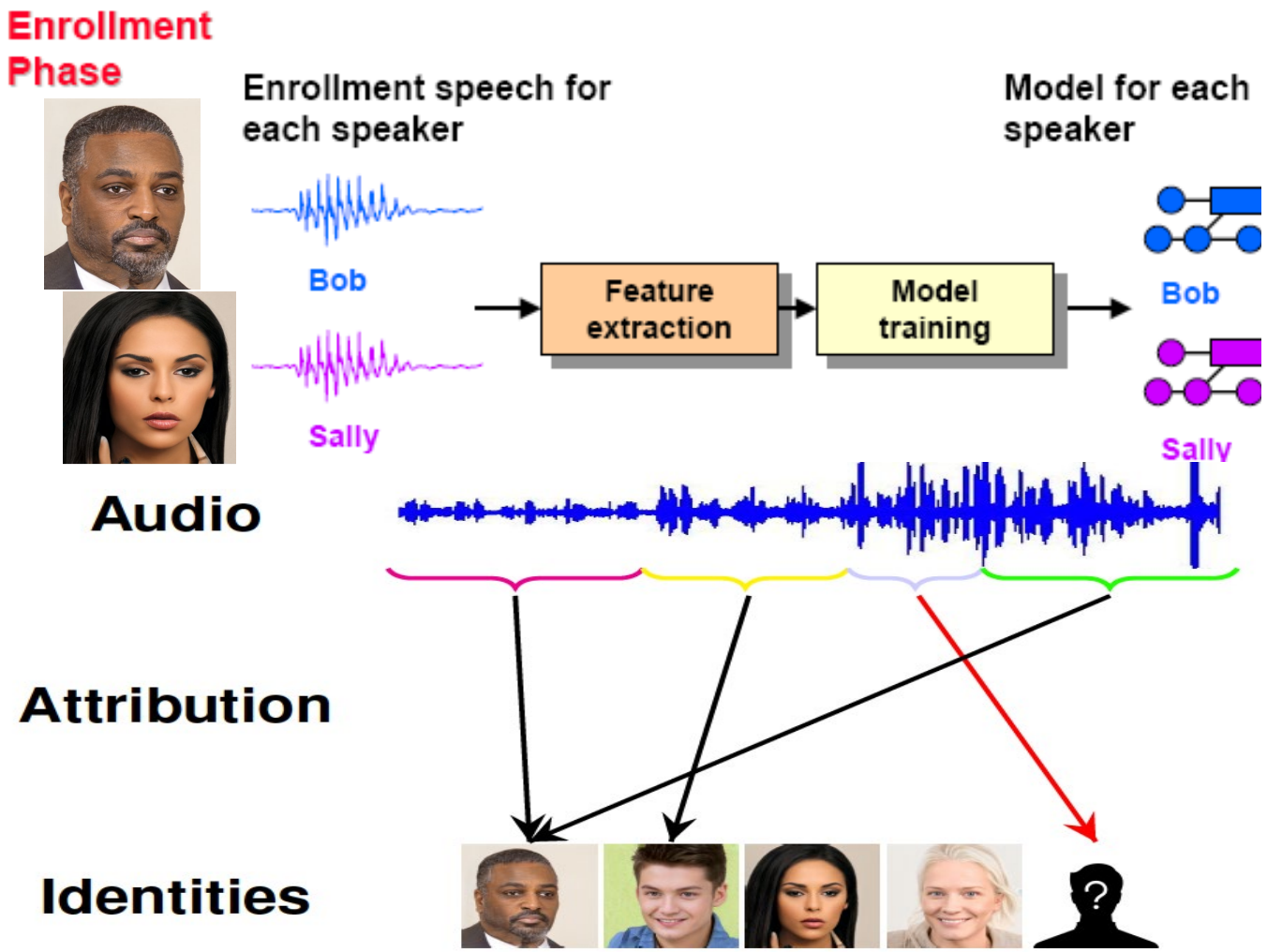
- **¿Para qué sirve?:**
  - Permite asignar identidades concretas a fragmentos de audio de un contenido analizado

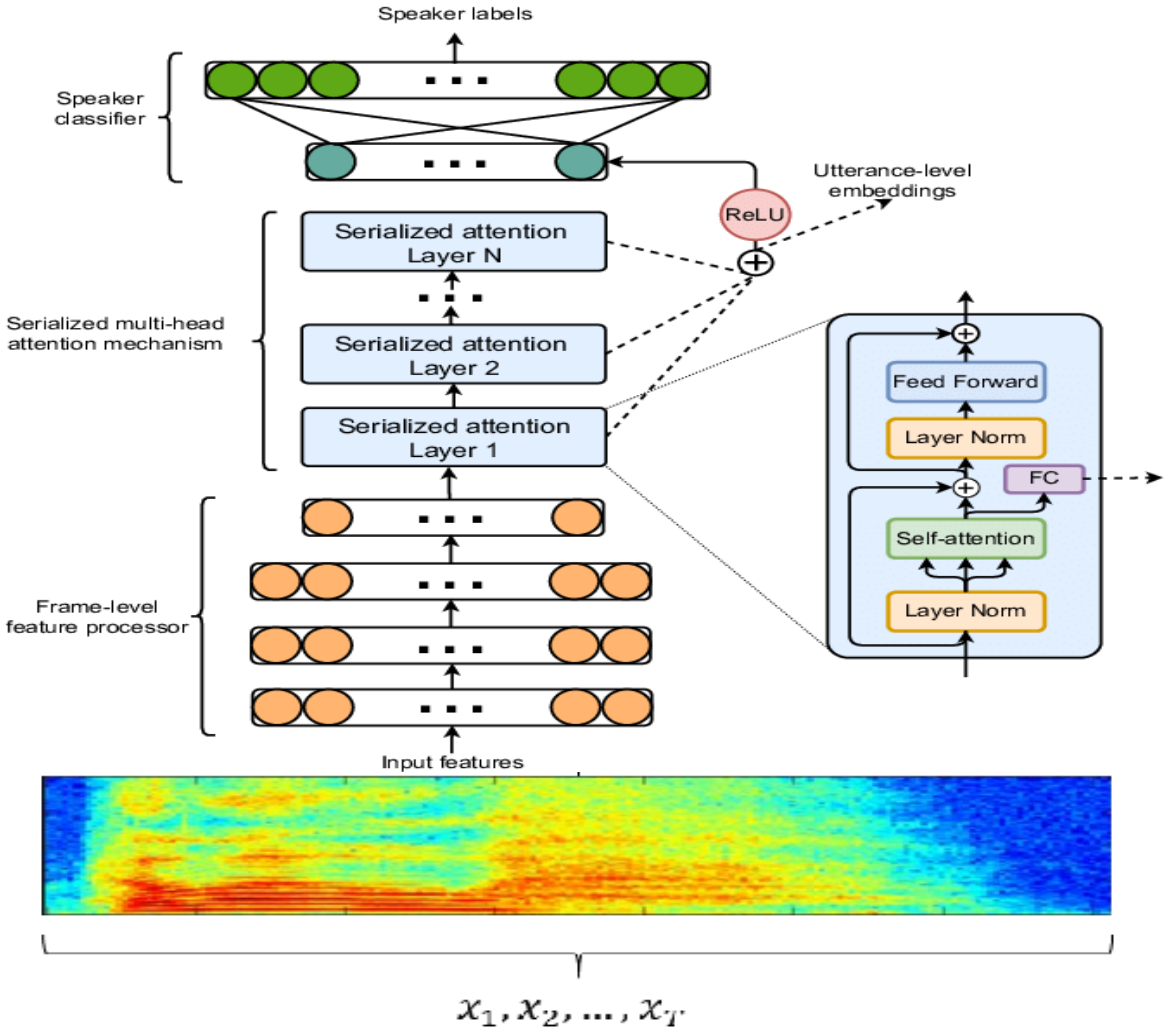
Retos en la Identificación de hablantes

- **Alta variabilidad en la voz de los hablantes**
- **Diversidad de dominios acústicos**
- **Solape entre hablantes**



Identificación de Hablantes





Viñals, I.; Ortega, A.; Miguel, A.; Lleida, E. **The Domain Mismatch Problem in the Broadcast Speaker Attribution Task** Applied Sciences, vol. 11, no. 18, pp. 8521, 2021.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). **Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification**. arXiv preprint arXiv:2005.07143.

VIVOLAB

Universidad Zaragoza

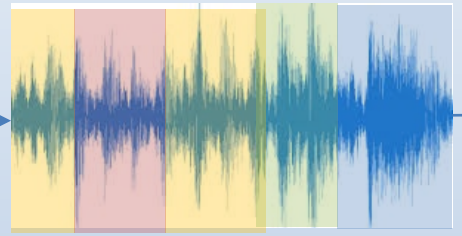
Identificación de Hablantes

- Reconocimiento de personajes en programas de TV:

Extract Audio and Content

Diarize Audio

Spk A Spk A Spk D



Spk B Spk C

Assign Speaker Identities

Speaker Verification Engine

Enroll Speakers

SPK Database

Spk Metadata File

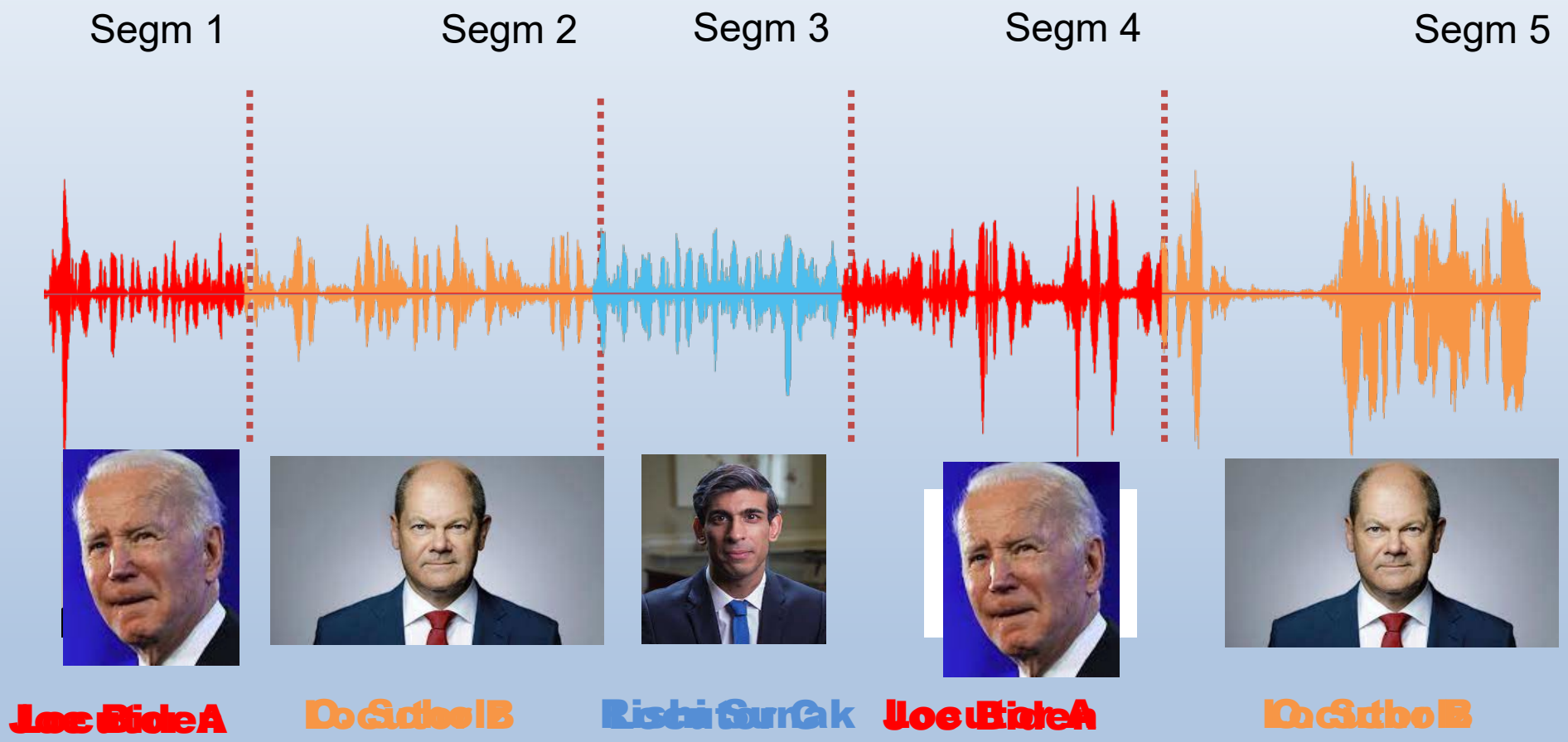
FileA John Smith start1 end1
File A John Smith start2 end2
File B John Smith start1 end1
File C John Smith start 1 end1
FileB Mary Adams start1 end 1
File C Sam Sanders start1 end 1

...

VIVOLAB

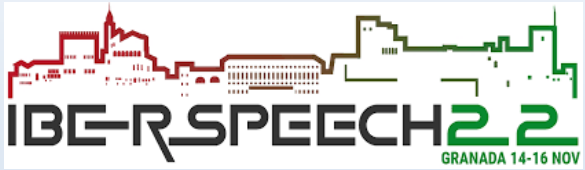
Universidad Zaragoza
1542

Diarización Junto con Identificación de Hablante:



Prestaciones: Albayzin – Retos RTVE

rtve Universidad Zaragoza 1542 RTTH

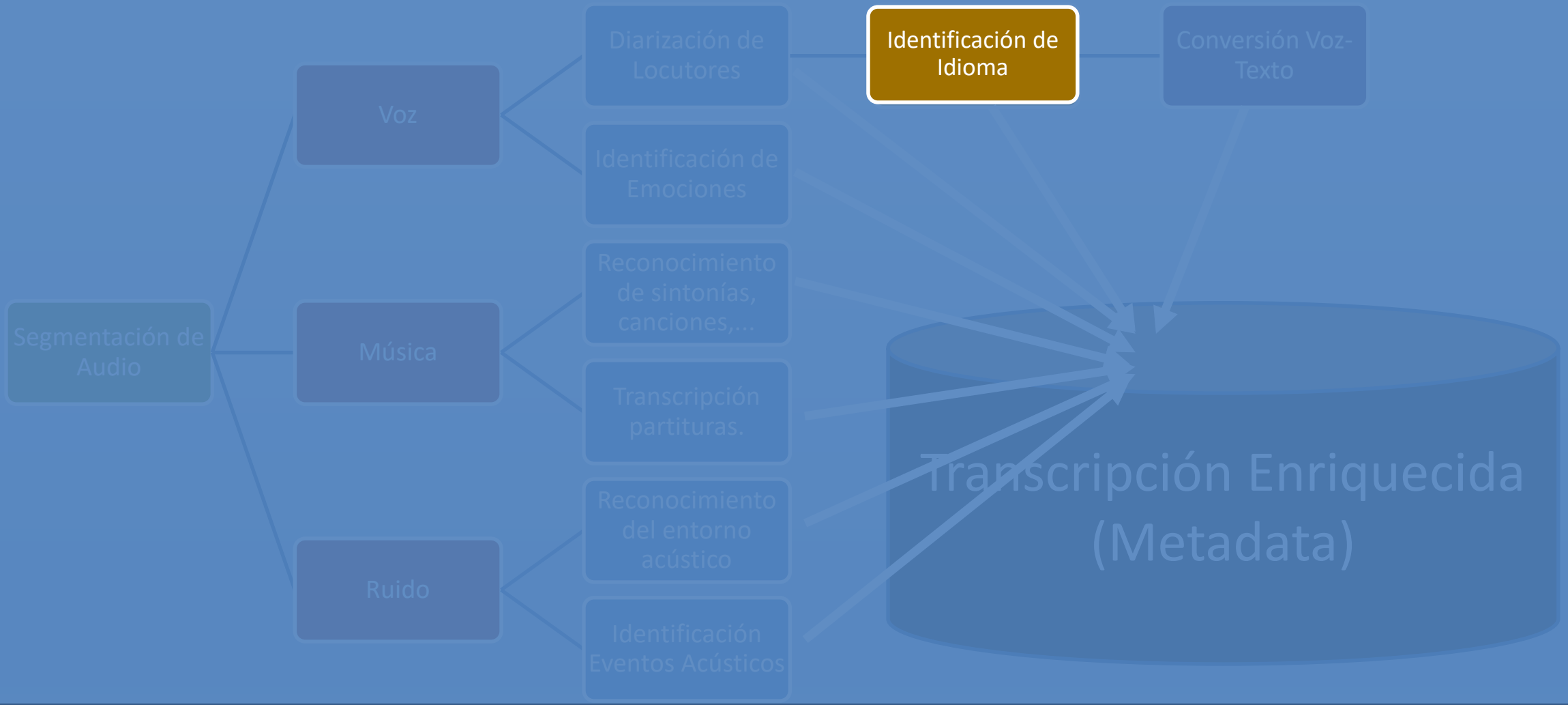


	AER
BIOMETRIC VOX	65.09 %
VIVOLAB	72.63 %

	MISS	FA	SPKERR	AER
TEAMIV_p	1,3	229,7	9,5	240,55
TEAMIV_l3	3,7	160,8	20,9	185,42
TEAMIV_l6	8,3	76,5	5,9	90,62
TEAMIV_l9	**12,4**	**15,3**	**1,2**	**28,88**

Subset	Closed Condition			Open Condition		
	Direct	Indirect	Hybrid	Direct	Indirect	Hybrid
Dev. subset	13.73	15.27	15.89	41.91	**37.45**	37.68
Eval. subset	25.11	17.20	**16.49**	65.31	60.34	**31.95**

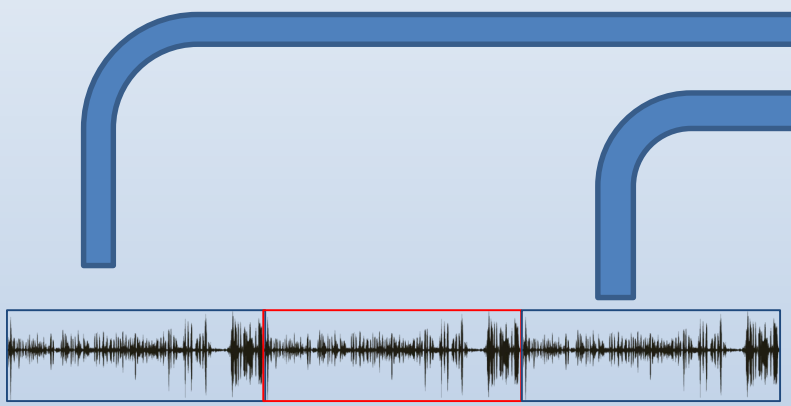
Identificación de Idioma



Identificación de Idioma

- **¿Para qué sirve?:**
  - **En entornos multilingües, permite el indexado y la recuperación de documentos:**
  - **Esencial en esos entornos como soporte a:**
    - **Reconocimiento automático del habla**

Identificación de Idioma



Conversión Voz-Texto (Castellano)

Conversión Voz-Texto (Catalán)

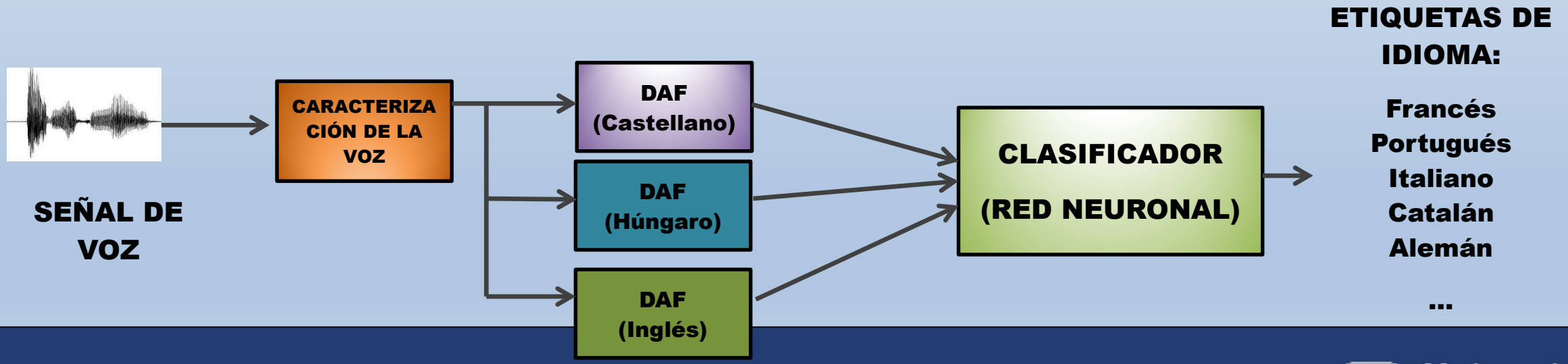
La Fundación ”la Caixa” ha hecho balance de la asistencia a sus centros y exposiciones culturales y científicas durante 2021, año todavía marcado por la pandemia. La entidad ha hecho un gran esfuerzo para recuperar el ritmo de visitantes en los ocho centros CaixaForum y en CosmoCaixa, y en las exposiciones itinerantes, para conseguir que la cultura sea segura siguiendo las restricciones de capacidad para el cóvido estipuladas por las comunidades autónomas. El público ha respondido y 5,2 millones de personas han visitado los centros y han vivido las diferentes propuestas culturales, científicas y educativas. Son 1,3 millones de personas más que en 2020 (+35,7%).

Així, set de cada deu visitants habituals abans de la pandèmia han tornat als centres CaixaForum i CosmoCaixa, un total d’1,9 milions de persones. Cal sumar-hi els 3,3 milions de visitants de les exposicions itinerants en més de 70 ciutats de tot Espanya.

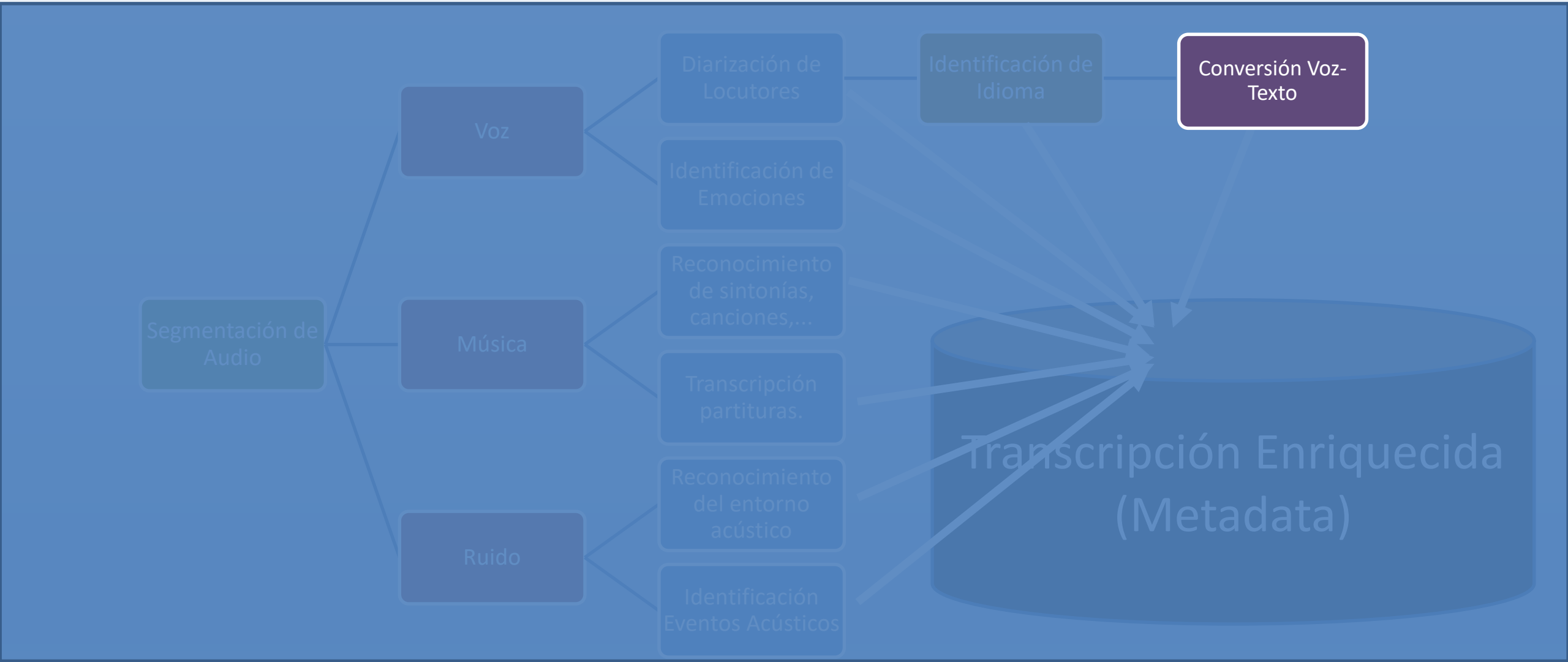
CaixaForum Barcelona, la més visitada, ha sido la exposición interactiva e inmersiva Teamlab. Arte, tecnología y naturaleza (en la imagen), seguida de El sueño americano. Del pop a la actualidad, con obras de Andy Warhol o Roy Lichtenstein, producida con la colaboración del British Museum. Por su parte, CosmoCaixa ha acogido, con gran afluencia de público, PRINT3D. Reimprimir la realidad y Nikola Tesla. El genio de la electricidad moderna, esta última vigente hasta febrero de 2022.

Identificación de Idioma

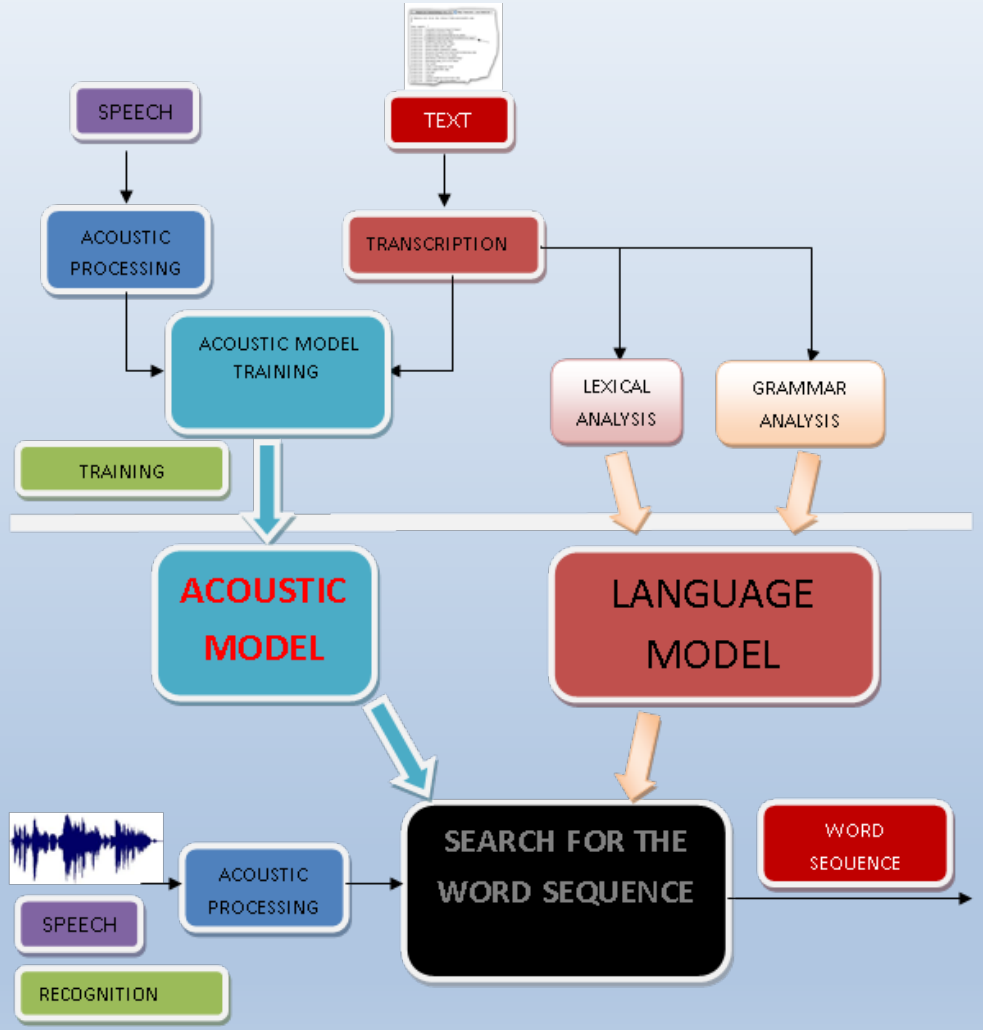
- **Acústicos**
  - Tratan de buscar patrones discriminativos directamente sobre la señal de voz
- **Fonotácticos (Lingüísticos)**
  - Primero procesan la señal de entrada con un (o varios) reconocedor fonético (en varios idiomas) después buscan patrones discriminativos en las secuencias de fonemas de salida

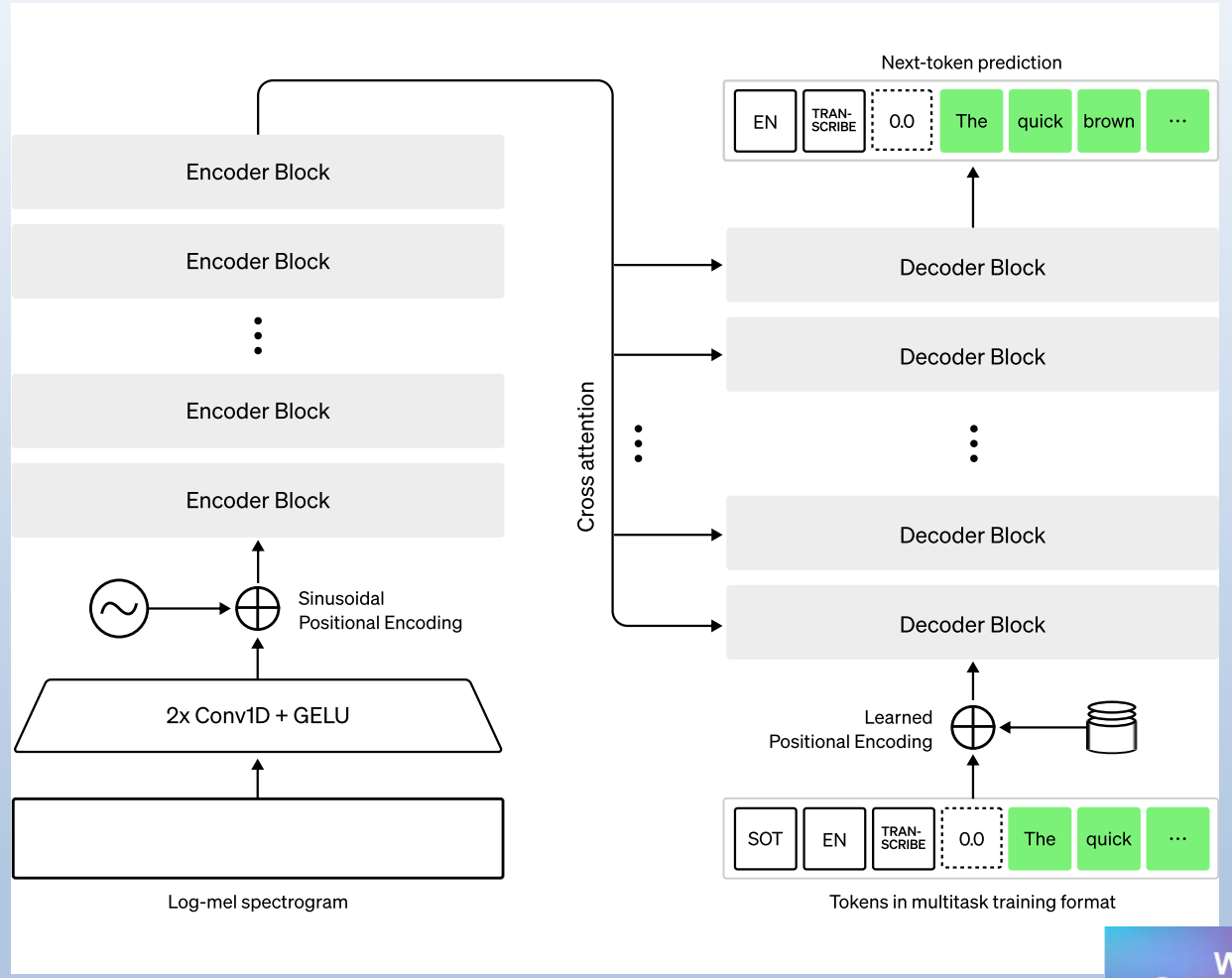


Reconocimiento Automático del Habla



Etapas y Procesos RAH:





https://openai.com/research/whisper

WHISPER OpenAI

Componentes de un Sistema de RAH

- **MODELO ACÚSTICO**
  - Describe las características de cada unidad desde el punto de vista de la señal de voz (espectralmente)
- **MODELO DE LENGUAJE**
  - Describe las relaciones entre palabras del vocabulario
  - Cuantifica la probabilidad de las secuencias de palabras
- **MODELO LÉXICO**
  - Describe cómo se forma cada palabra del vocabulario a partir de las diferentes unidades del modelo acústico.

Errores en un Sistema de RAH

- **Borrados**
  - El locutor dice algo pero el sistema no devuelve nada
- **Substituciones**
  - El sistema devuelve a su salida una palabra diferente de la pronunciada por el locutor.
- **Inserciones**
  - El locutor no dice nada, pero el sistema devuelve alguna palabra (generalmente debido a artefactos acústicos)

Errores en un Sistema de RAH

- Métricas de Precisión y Error:

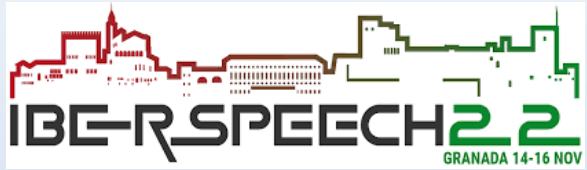
REF: a las tres y siete minutos de mañana
HYP: a las tres diecisiete minutos de la mañana

CORRECTO (C)
ERRORES:
Substitutiones (S), Borrados (B), Inserciones (I)

$$\%\ ACC = \frac{C}{C+S+B+I} \times 100$$

$$\%\ WER = \frac{S+B+I}{C+S+B} \times 100$$

Prestaciones: Albayzin – Retos RTVE



Show	Hours
3x4	2:58:17
A pedir de boca (APB)	3:41:38
Agrosfera (AG)	4:15:20
Aquí la Tierra (AT)	2:46:44
Ateneo (ATE)	1:40:09
Cerámica Popular Española (CPE)	1:02:35
Comando Actualidad (CA)	3:59:29
Conversatorios Casa América (CCA)	1:58:44
Corazón (CO)	3:00:17
El cazador (EC)	3:48:22
Encuestas ruido ambiente (ERA)	2:08:13
Entrevistas en bruto (EE)	3:54:57
España Directo (ED)	4:05:57
Fiction (Grasa-GR, Yrreal-YR, Riders_RD)	3:53:22
Informativos UMATIC (IU)	0:59:49
Jara y Sedal (JYS)	2:29:17
Noticias Nacional (NN)	2:14:32
Saber y Ganar (SYG)	4:28:28
Toros (TO)	0:49:57
21 different shows	54:16:07

Diarización e Identificación de hablantes, Añadiendo Reconocimiento del Habla:

J. Biden: Contenido de la intervención 1 … <Tcomienzo1> <Tfin1>

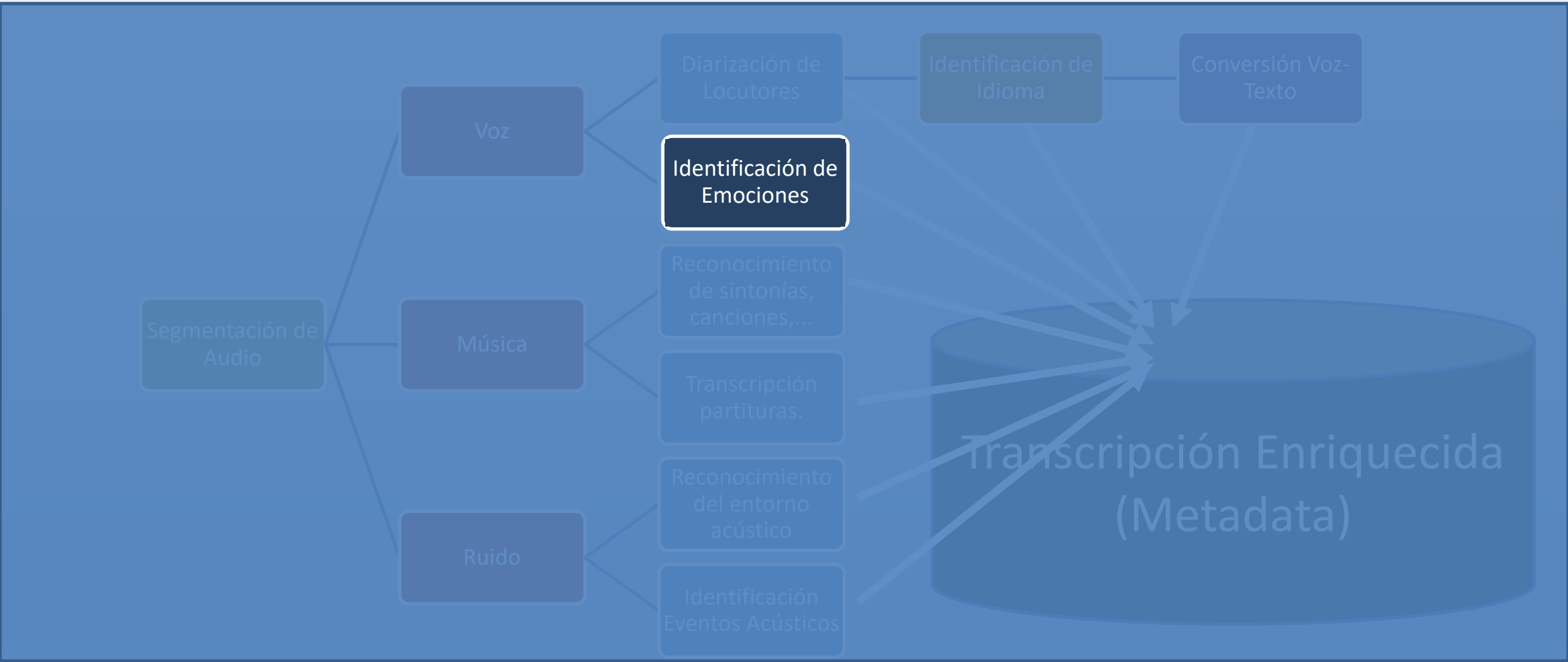
O. Scholz : Contenido de la intervención 2 … <Tcomienzo2> <Tfin2>

Rishi Sunak: Contenido de la intervención 3 … <Tcomienzo3> <Tfin3>

J. Biden: Contenido de la intervención 4 …. <Tcomienzo4> <Tfin4>

A. Scholz : Contenido de la intervención 5 … <Tcomienzo5> <Tfin5>

Tecnologías



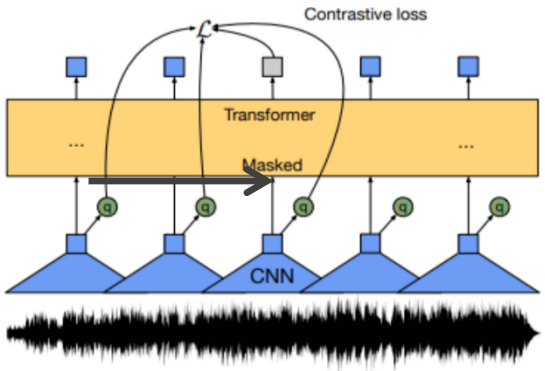
Identificación de Emociones

- **¿Para qué sirve?:**
  - **Puede añadir información extra que enriquece el discurso de los protagonistas de un contenido**

Identificación de Emociones

SEÑAL DE VOZ

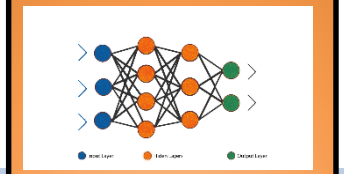
Feature Extraction Using Wav2Vec2



Classify Layer

Feature Vectors

CLASIFICACIÓN



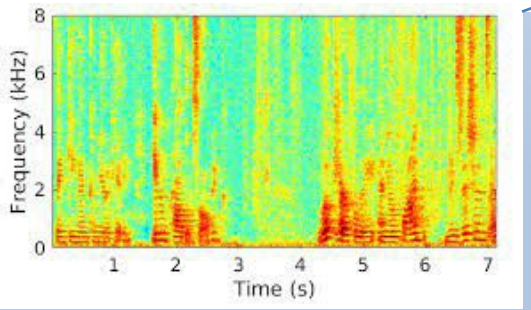
Emotion Recognition

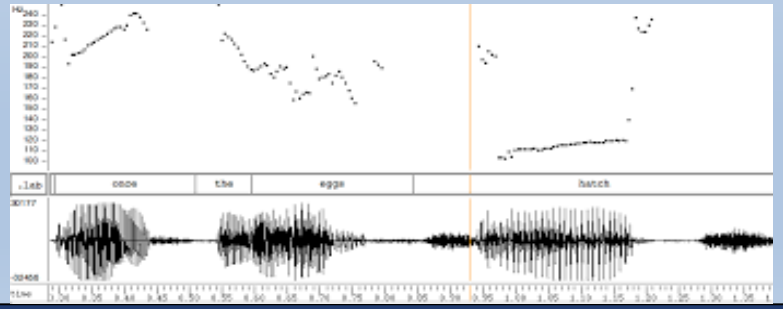
- enfado
- alegría
- tristeza
- neutro
- aburrimiento
- ansiedad

Características:

Espectrales
Prosódicas
Paralingüísticas

...





Síntesis de voz

https://github.com/coqui-ai/TTS

El último juguete

Bark

Universidad
Zaragoza