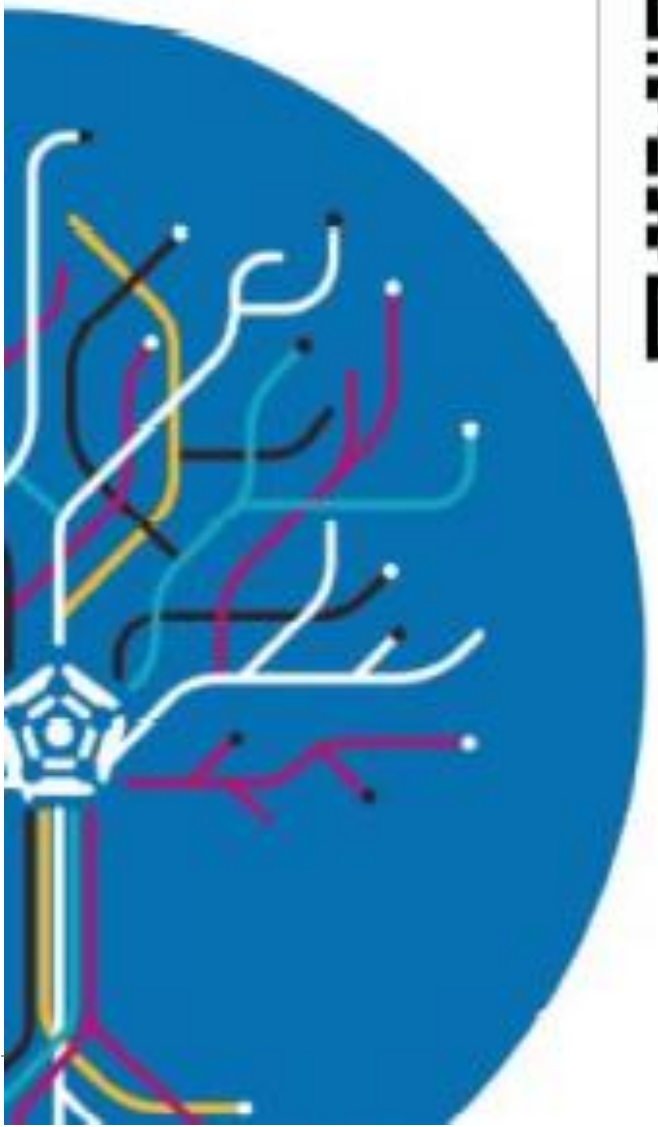
Universidad Zaragoza

1542

Curso 2025

CURSO

"¿QUÉ TIENE LA INTELIGENCIA ARTIFICIAL PARA MI ARCHIVO? DE LA TEORÍA A LA PRÁCTICA"



universidad zaragoza

VIVOLAB

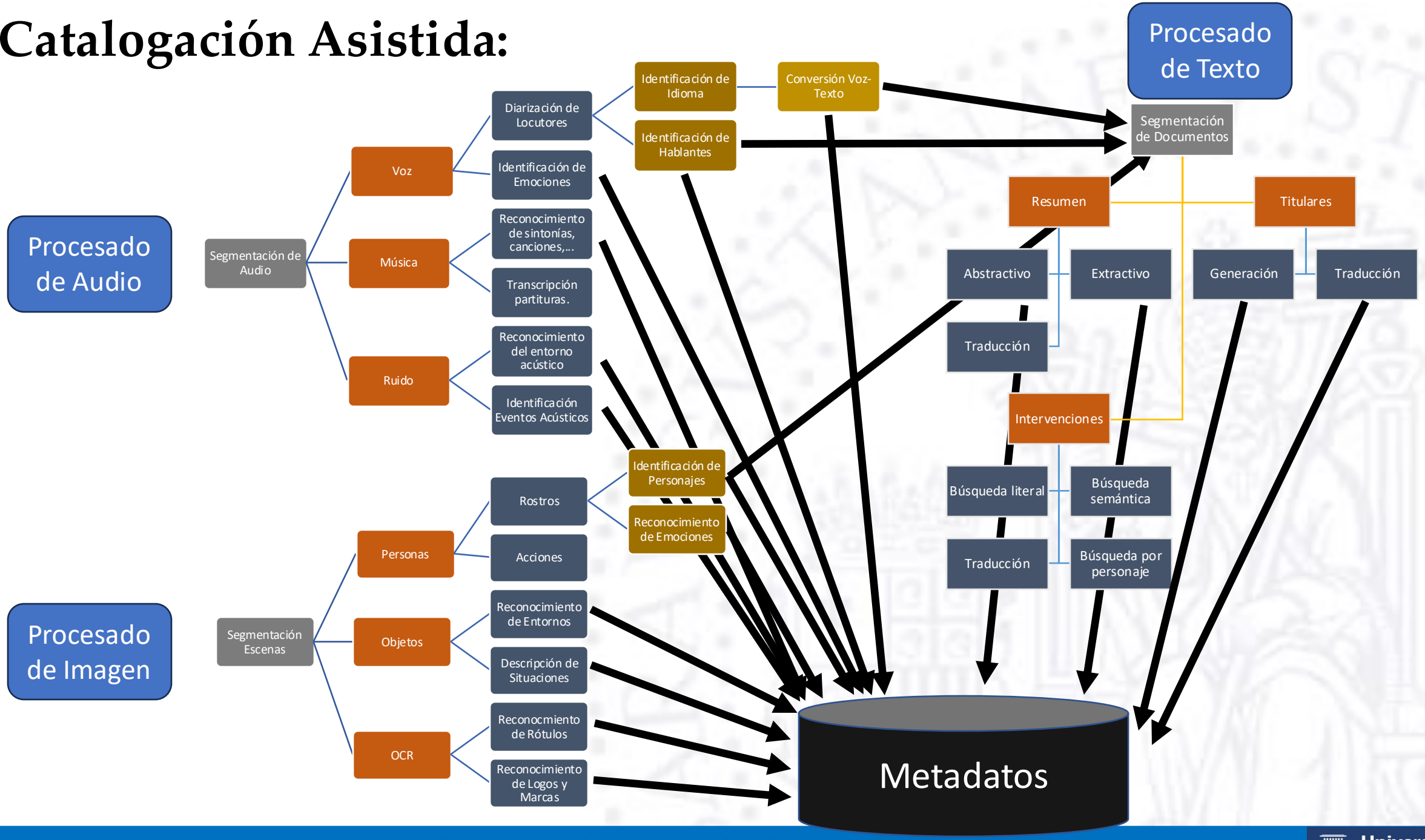
¿Qué tiene la IA para mi archivo?

Análisis y extracción de metadatos en contenidos audiovisuales:

Tecnologías del Habla

Universidad Zaragoza

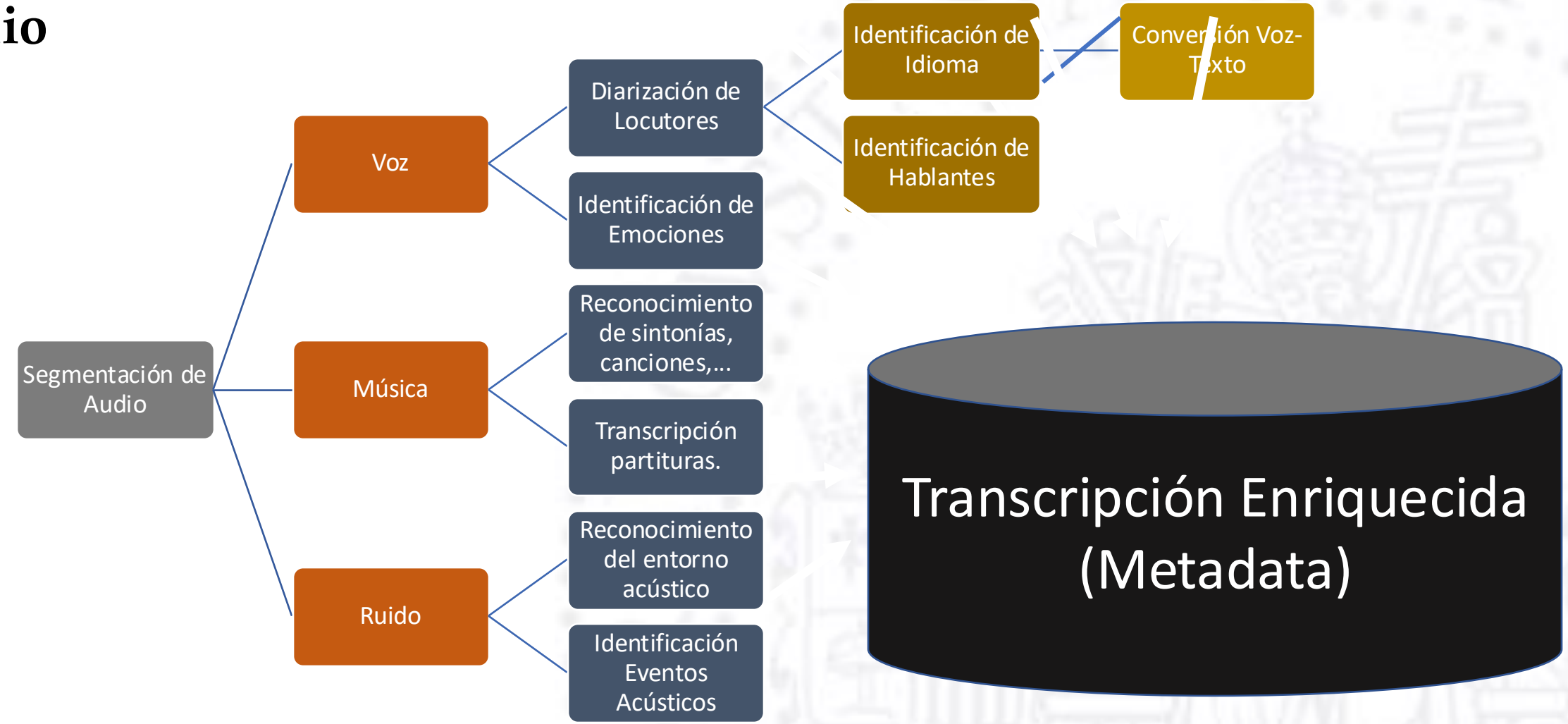
Catalogación Asistida:



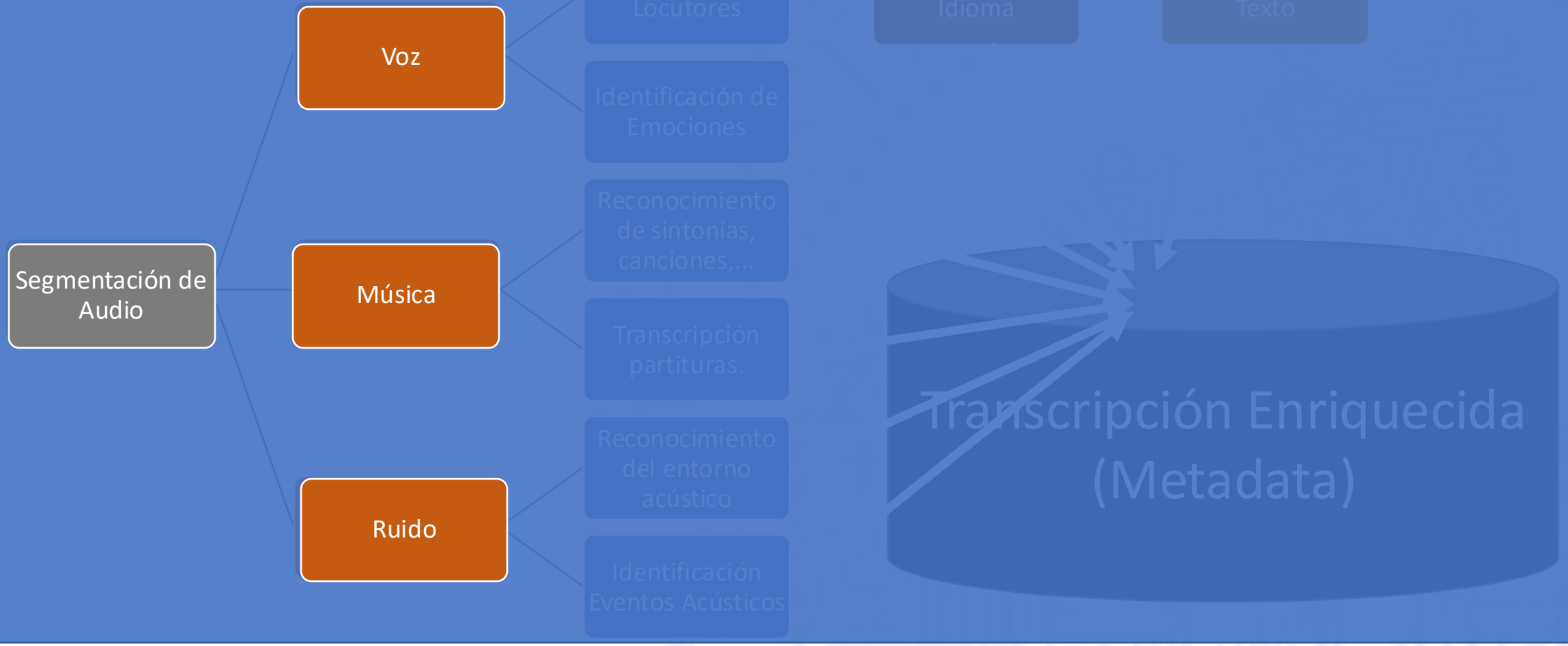
¿Qué información podemos encontrar en un audio?

- 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 Audio



Segmentación de Audio



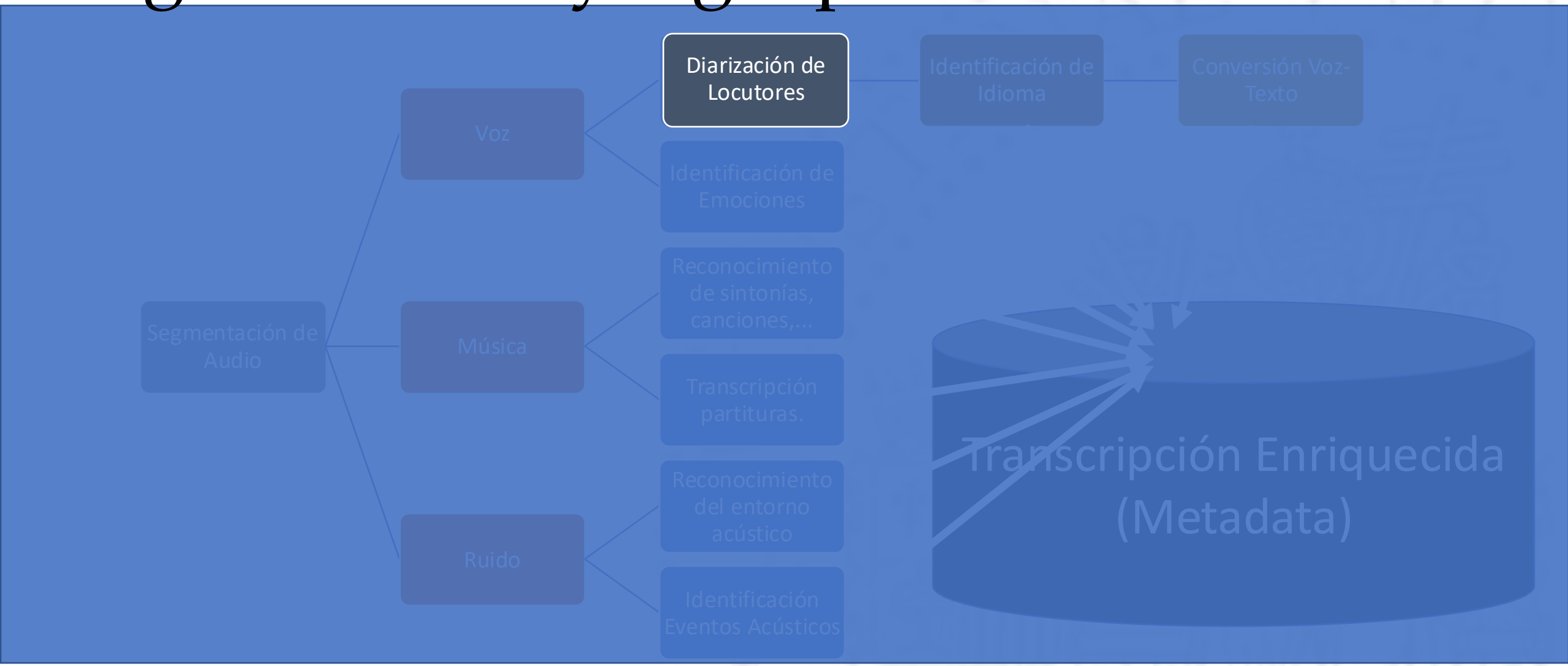
Segmentación de Audio

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



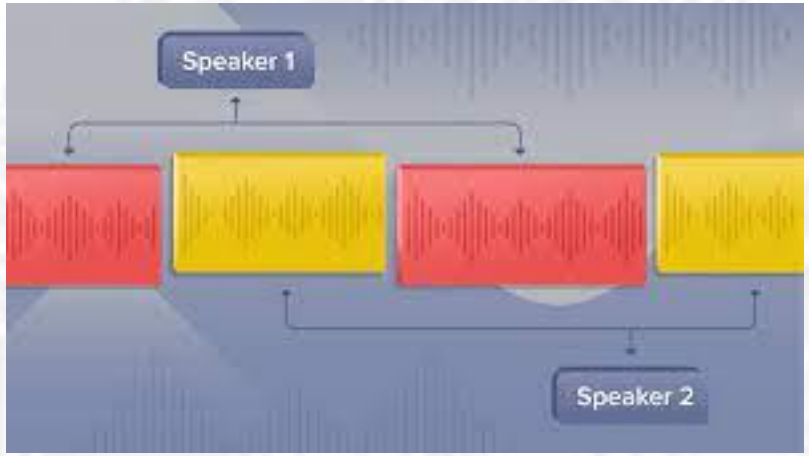
- **¿Para qué sirve?:**
  - **Da soporte a otras tareas de extracción de información como:**
    - **Diarización**
    - **Identificación del hablante**
    - **Conversión Voz-Texto**
    - **…**

Segmentación y Agrupación de Hablantes



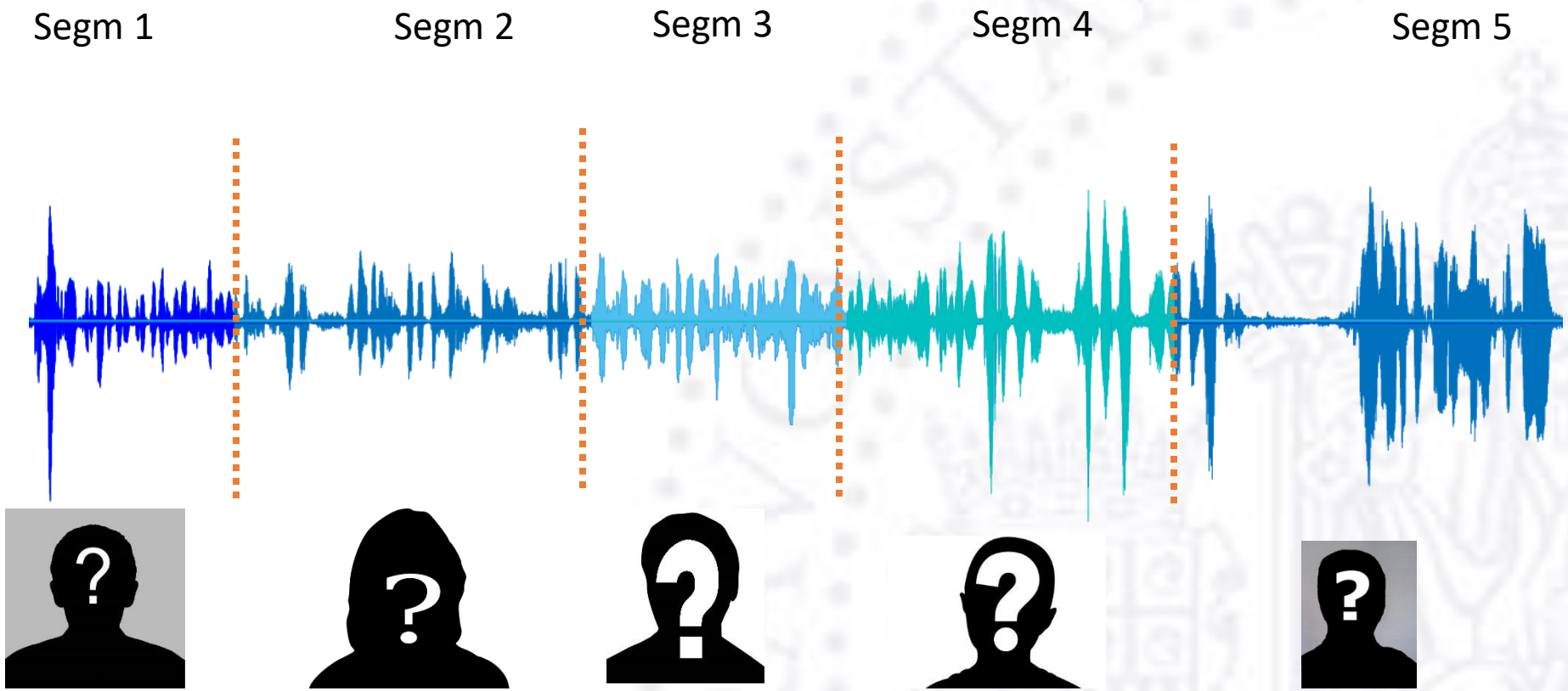
Segmentación y Agrupación de Hablantes

- **¿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 de:**
    - **Reconocimiento automático del habla**
    - **Reconocimiento del hablante**
    - **…**

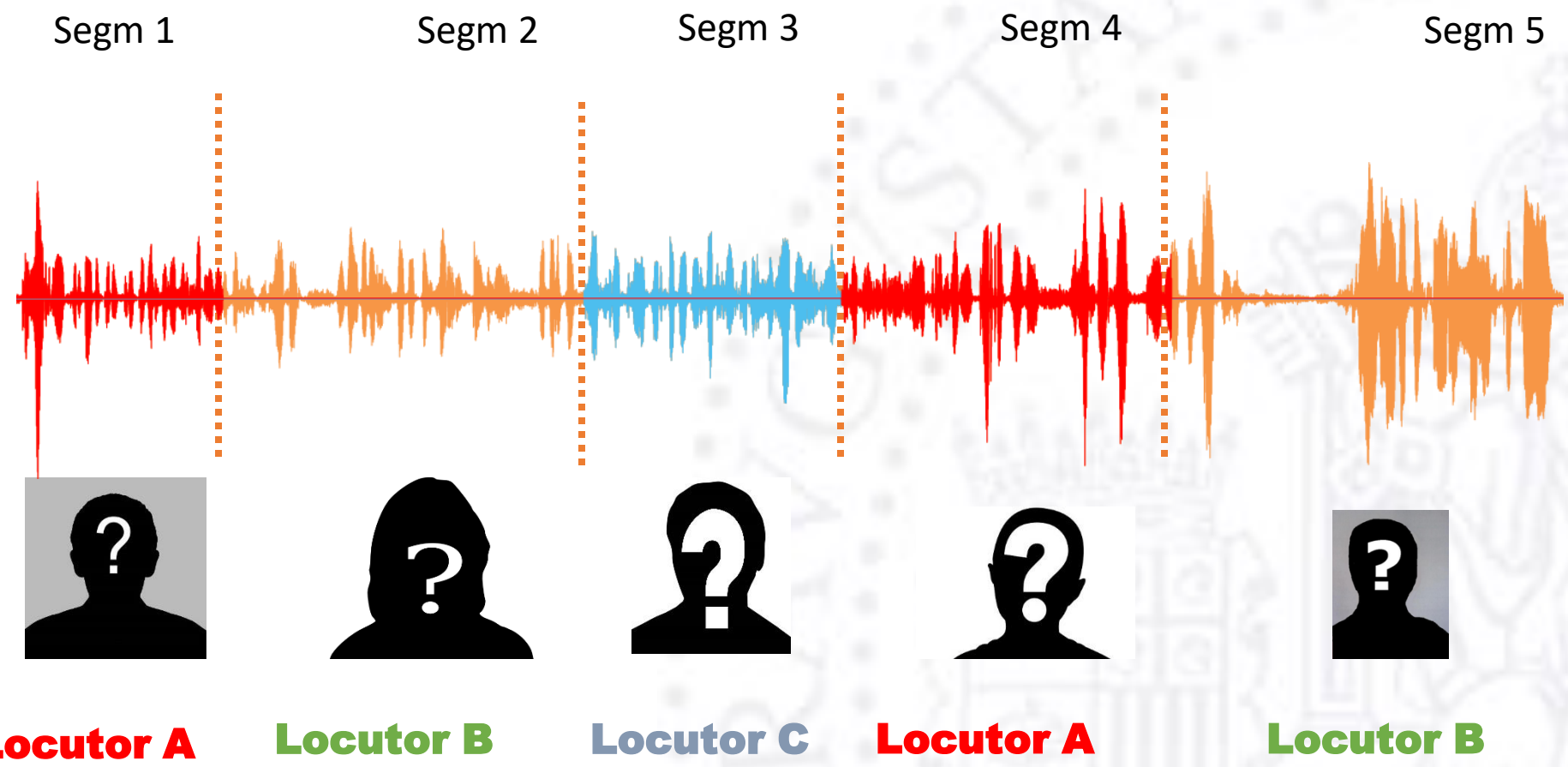


Segmentación y Agrupación de Hablantes

Segmentación y Agrupación de Hablantes

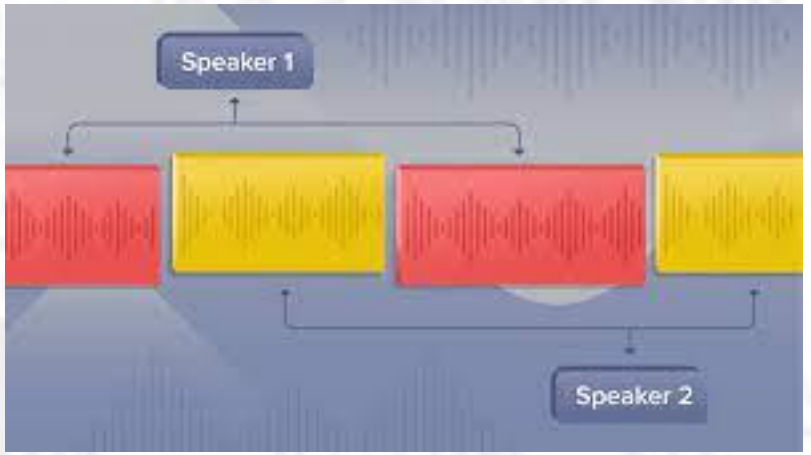


Segmentación y Agrupación de Hablantes



Segmentación y Agrupación de Hablantes

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

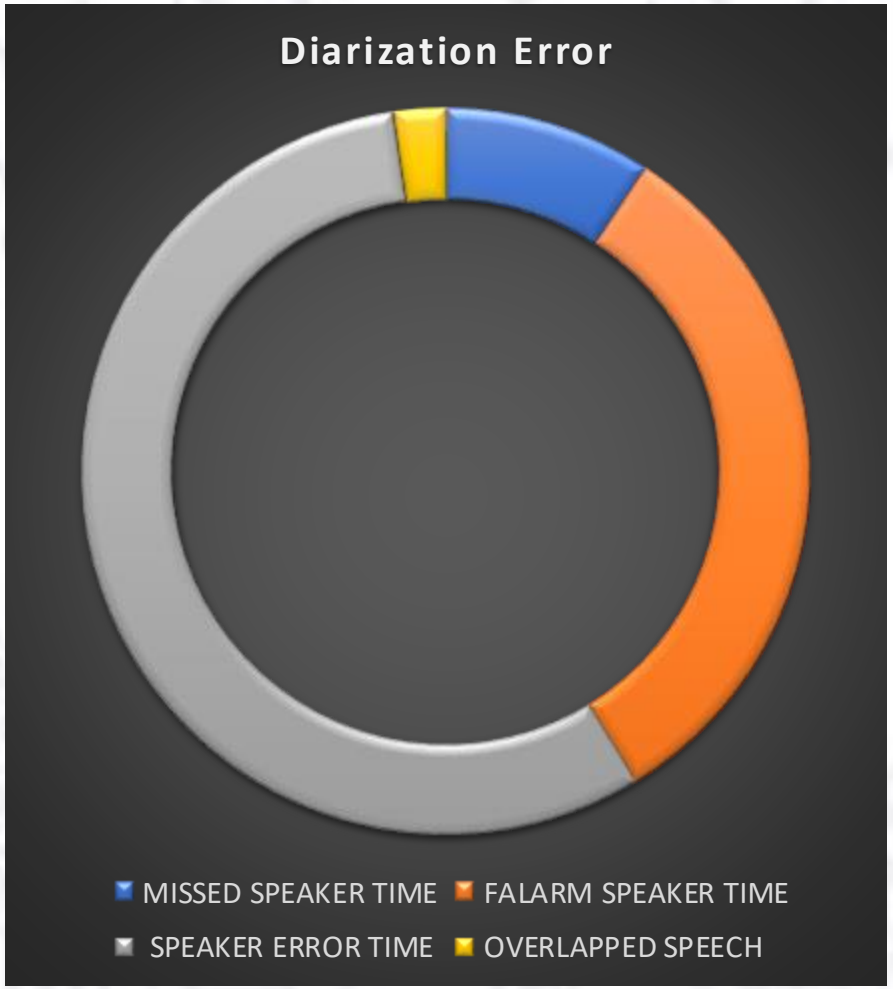


Medida del Error

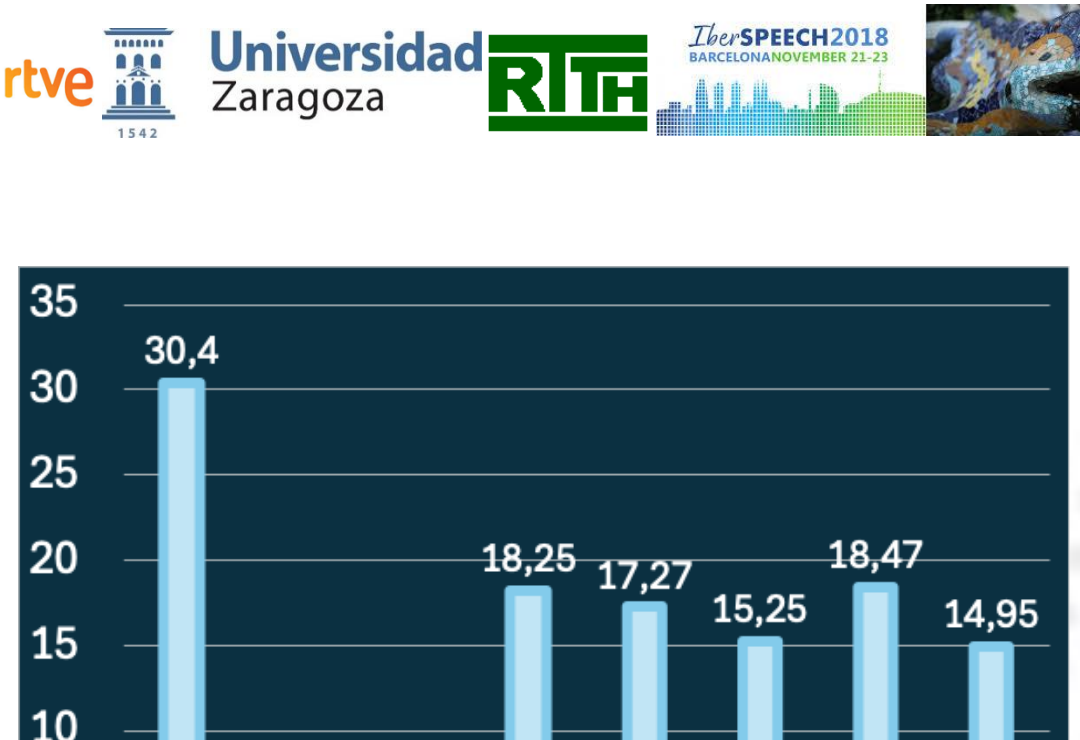
- **Diarization Error Rate:**

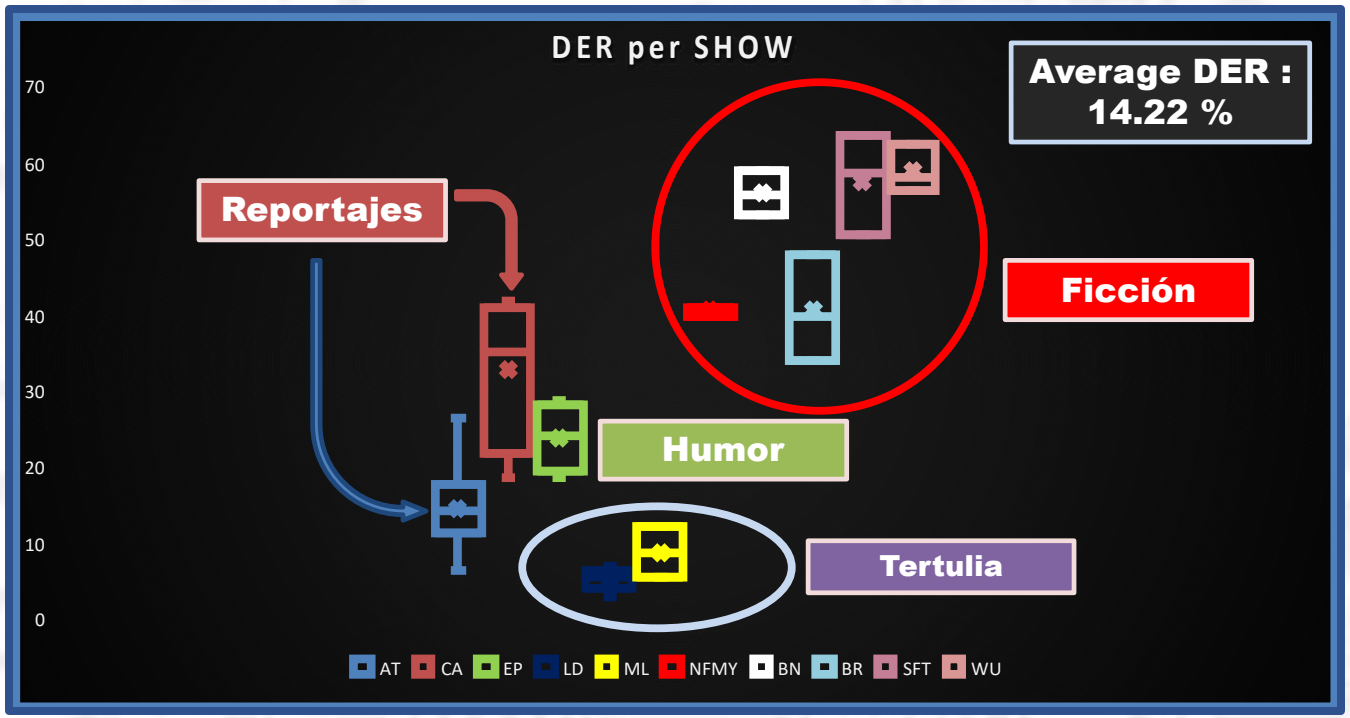
$$DER = \frac{T_{incorrecto}}{T_{voz}}$$

- **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.**

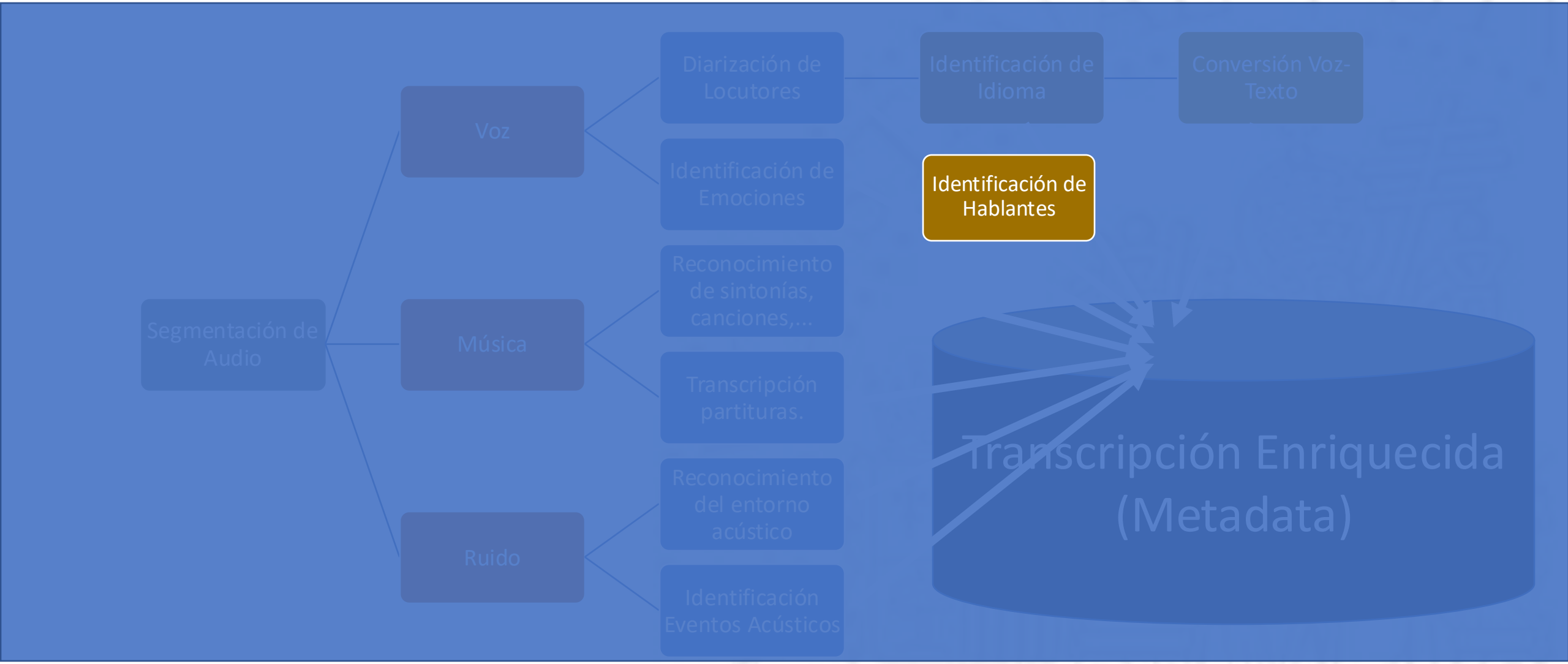


Prestaciones: Albayzín Retos - RTVE





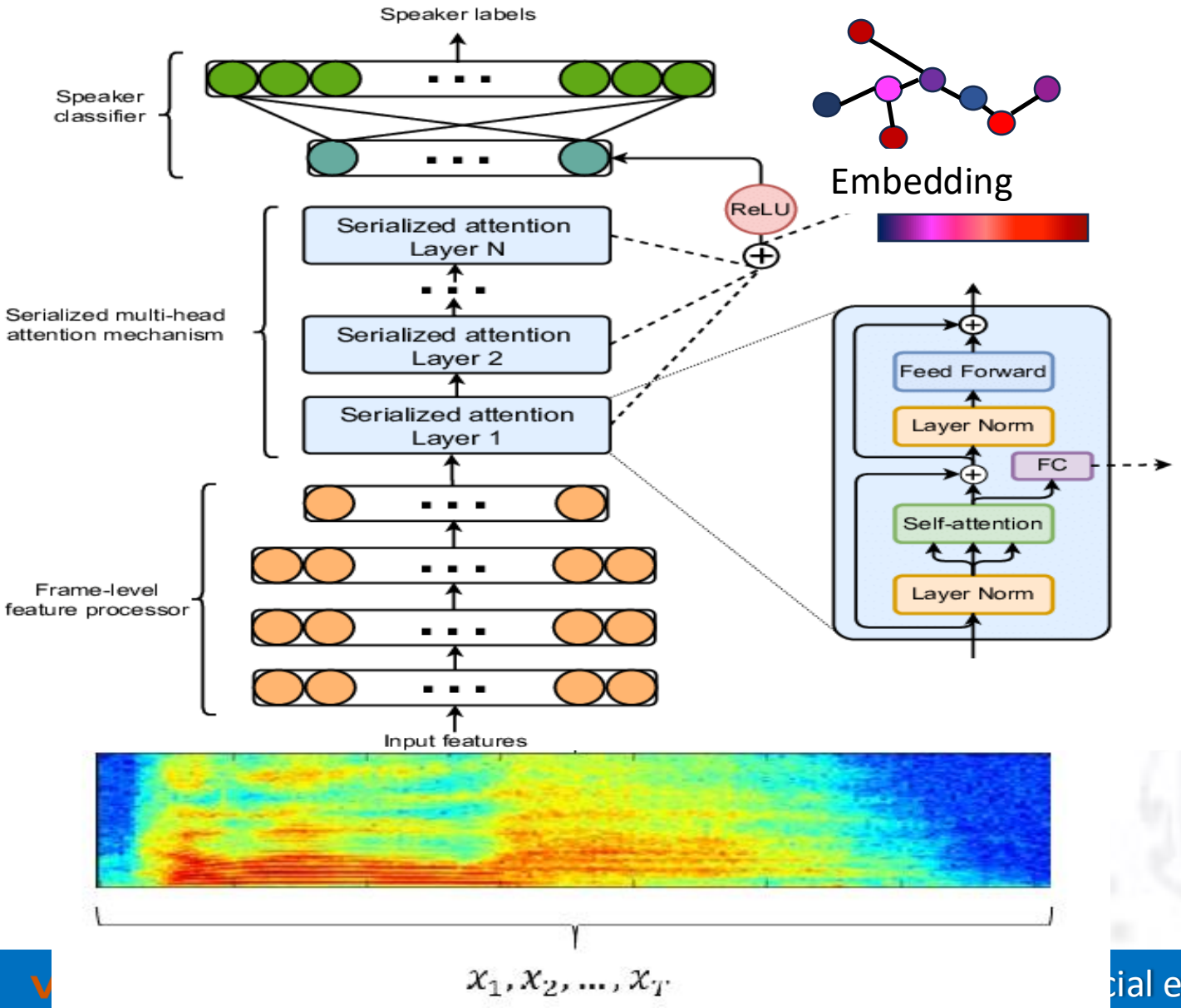
Identificación de Hablantes



Identificación de Hablantes

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

Cómo se hace …

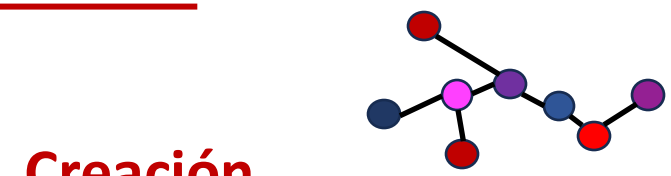


De cada fragmento de voz se extrae una "huella" que se compara con las almacenadas en la fase de registro para saber quién ha pronunciado ese contenido

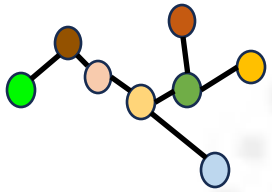
FASE DE REGISTRO

Bob

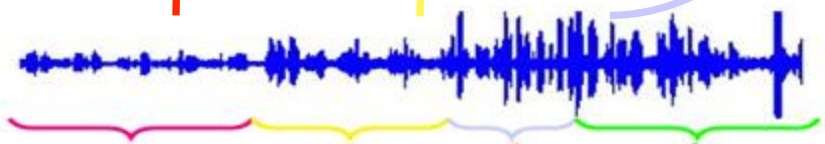
Creación de Modelos



Sally



AUDIO

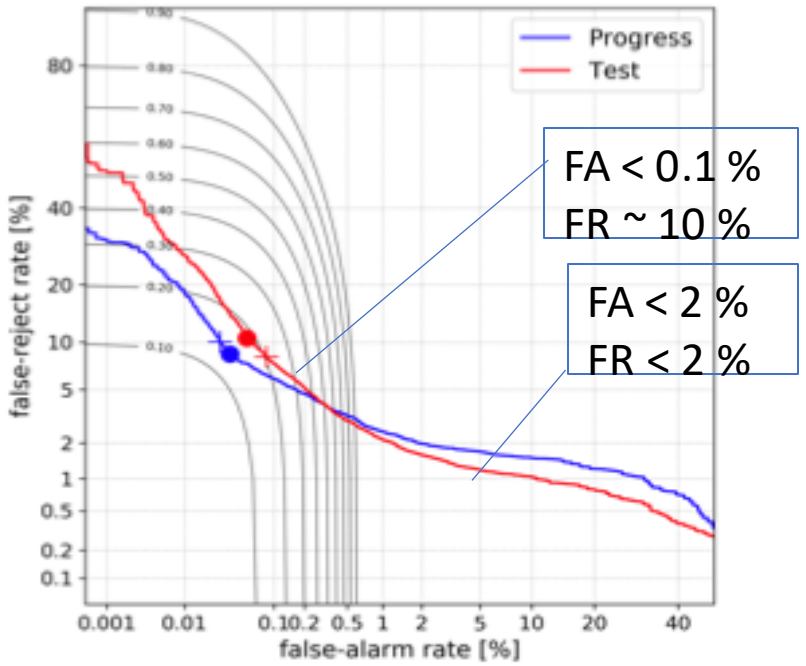


ATRIBUCIÓN

IDENTIDADES

FASE DE OPERACIÓN

Precisión ...



NIST
National Institute of Standards and Technology
U.S. Department of Commerce

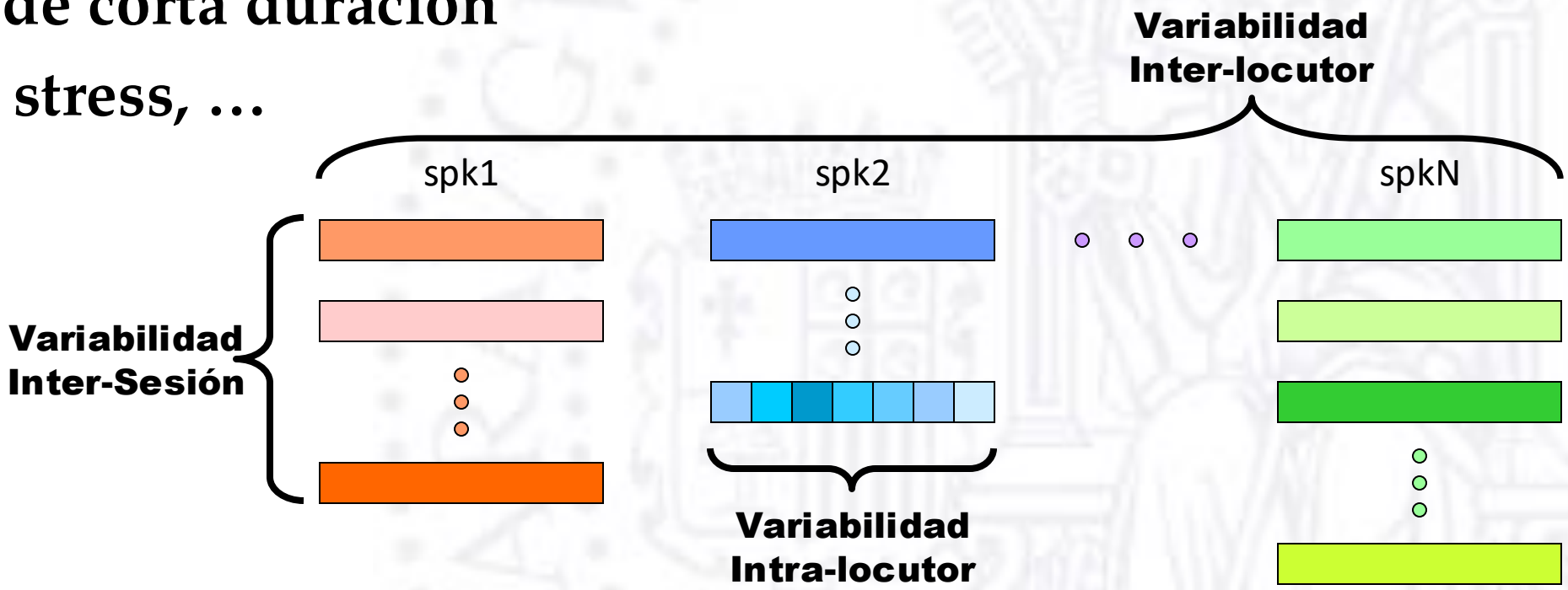
The 2019 NIST Speaker Recognition Evaluation CTS Challenge

Seyed Omid Sadjadi[1], Craig Greenberg[1], Elliot Singer[2,†], Douglas Reynolds[2,†], Lisa Mason[3], Jaime Hernandez-Cordero[3]

Redes Neuronales con cientos de millones de parámetros consiguen tasas de error por debajo del 2%

En entornos reales de operación …

- Alta variabilidad en la voz de los hablantes
- Diversidad de dominios acústicos (Estudio, Calle, Estadio, …)
- Solape entre hablantes
- Intervenciones de corta duración
- Voz emocional, stress, …

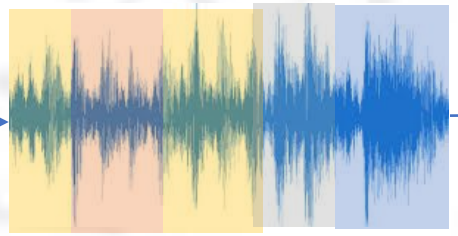


Identificación de Hablantes

- Reconocimiento de personas en programas de TV:

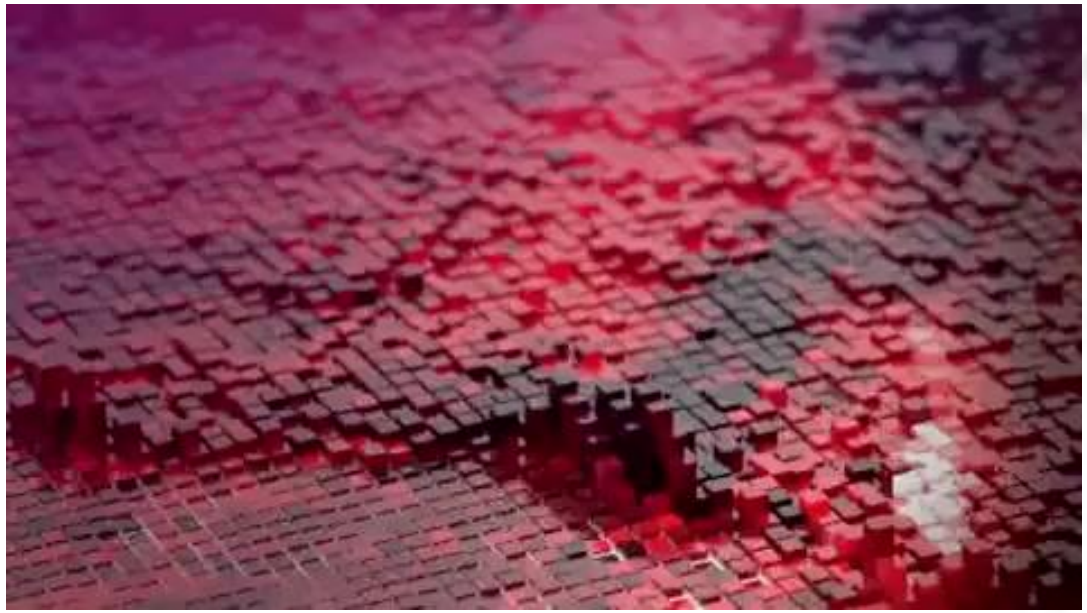
Extraer Audio

Diarizar Audio



Asignar Identidad Hablante

Motor Reconocimiento Biométrico



Registrar Locutores

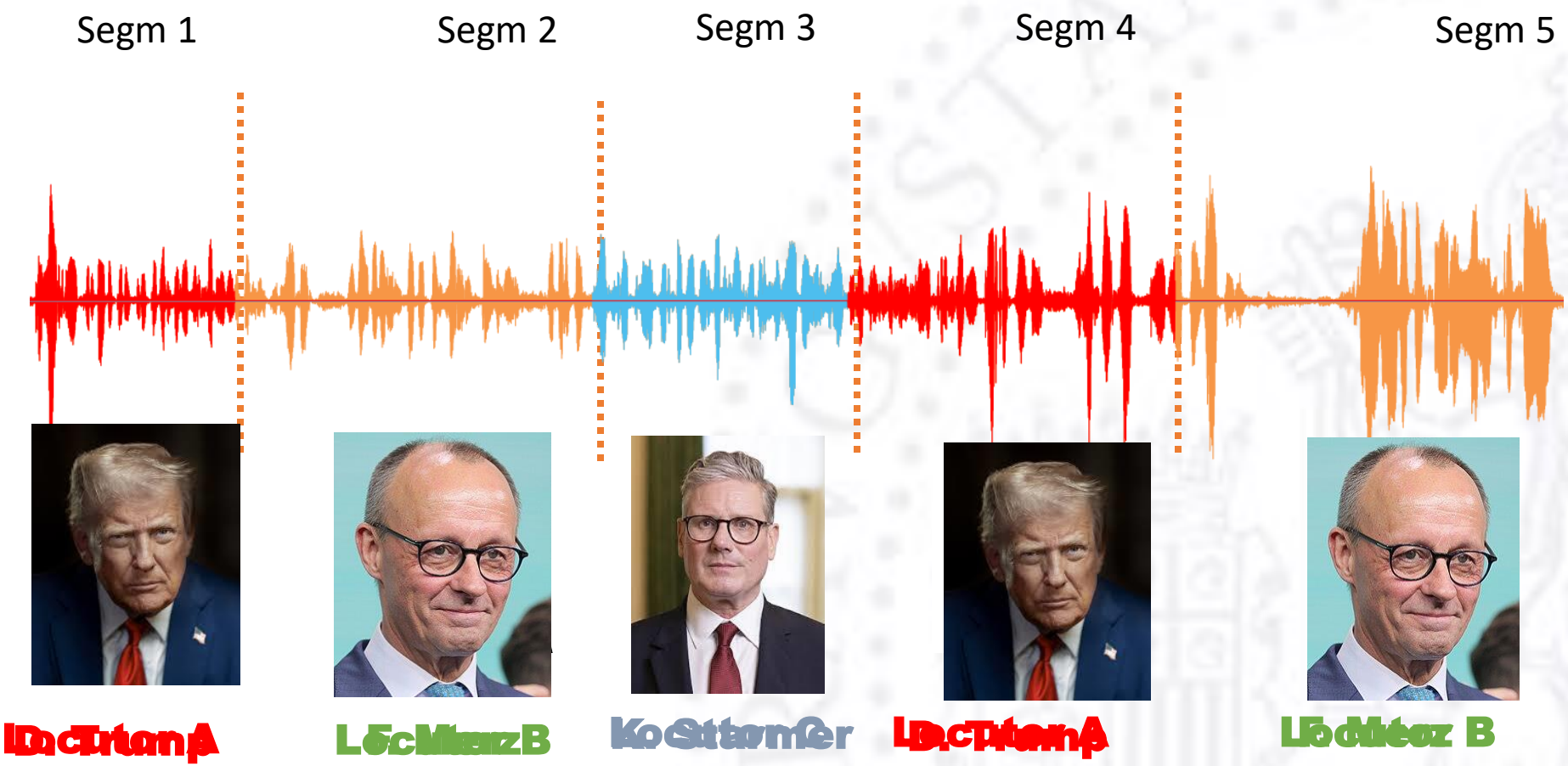
Base de Datos Hablantes

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

...

Diarización Junto con Identificación de Hablante:



Prestaciones: Albayzin – Retos RTVE

	AER
BIOMETRIC VOX	65.09 %
VIVOLAB	72.63 %

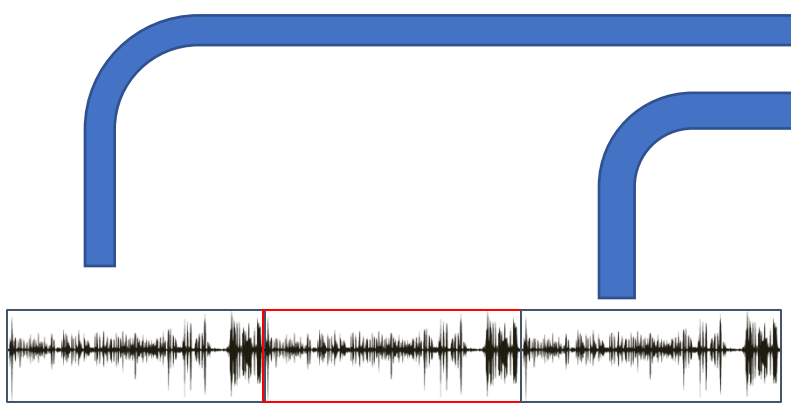
	MISS	FA	SPKERR	AER
TEAMIV_p	1,3	229,7	9,5	240,55
TEAMIV_l3	3,7	160,8	20,9	185,42
TEAMIV_l6	8,3	76,5	5,9	90,62
TEAMIV_l9	**12,4**	**15,3**	**1,2**	**28,88**

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

- **¿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)

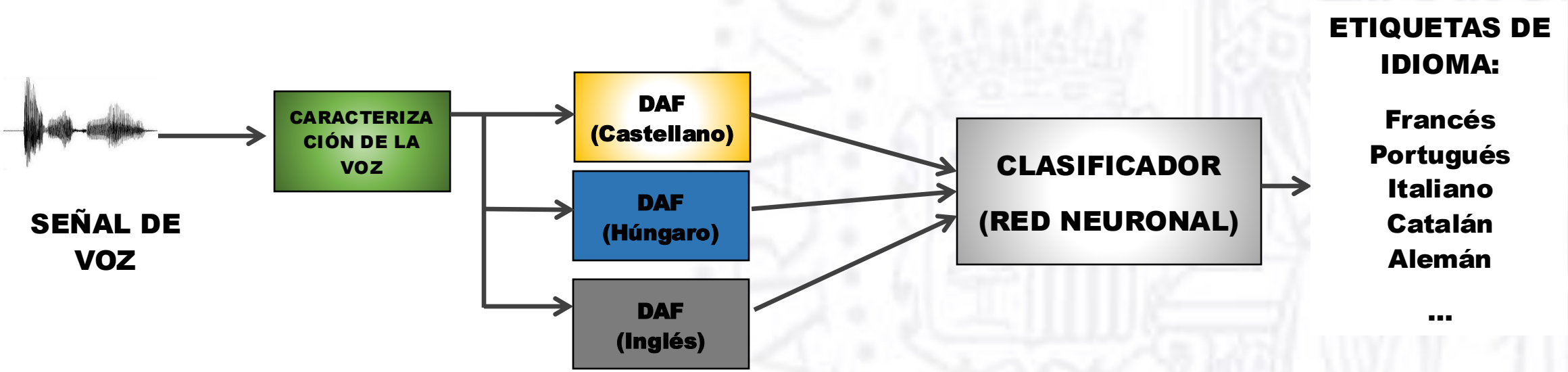
La Fundación ”la Caixa” ha hecho balance de la asistencia a sus centros y exposiciones culturales y científicas durante 2021, año todavía marcado por la pandemia. La entidad ha hecho un gran esfuerzo para recuperar el ritmo de visitantes en los ocho centros CaixaForum y en CosmoCaixa, y en las exposiciones 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.

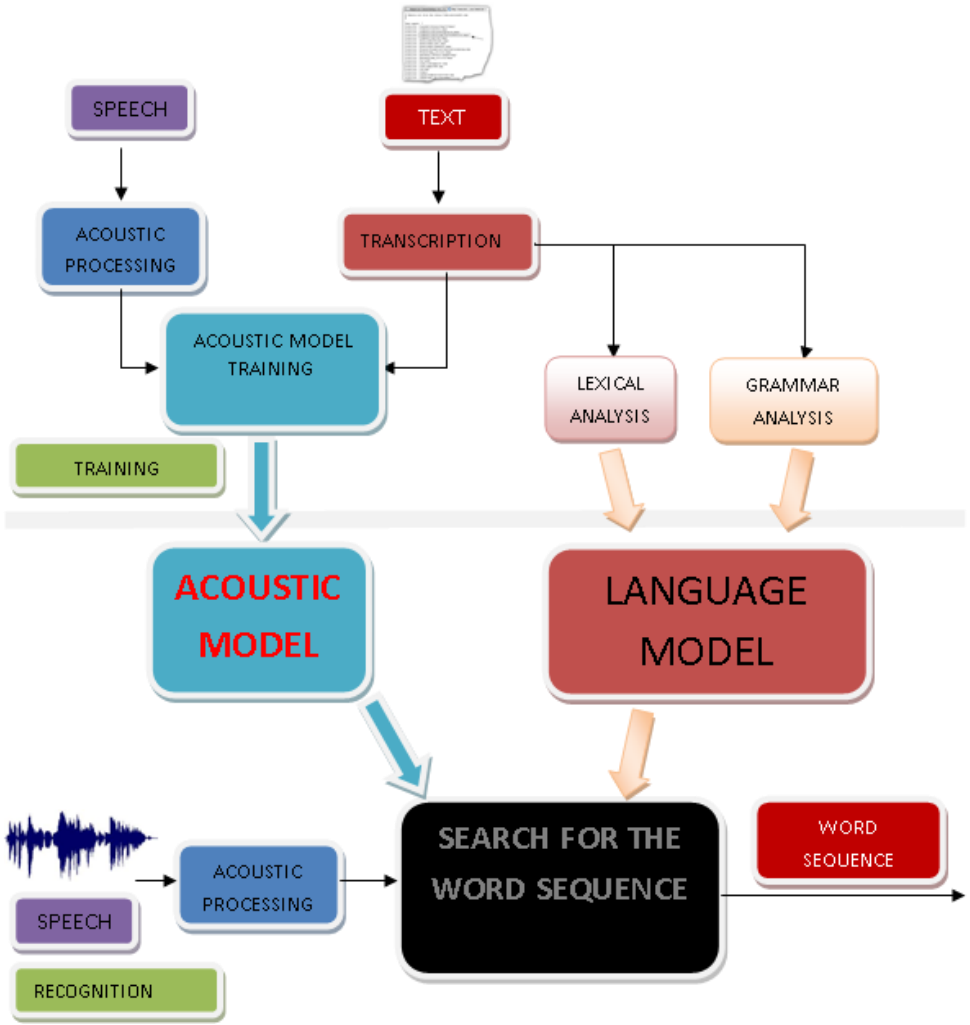
CaixaForum Barcelona, la más visitada, ha sido la exposición interactiva e inmersiva Teamlab. Arte, tecnología y naturaleza (en la imagen), seguida de El sueño americano. Del pop a la actualidad, con obras de Andy Warhol o Roy Lichtenstein, producida con la colaboración del British Museum. Por su parte, 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.

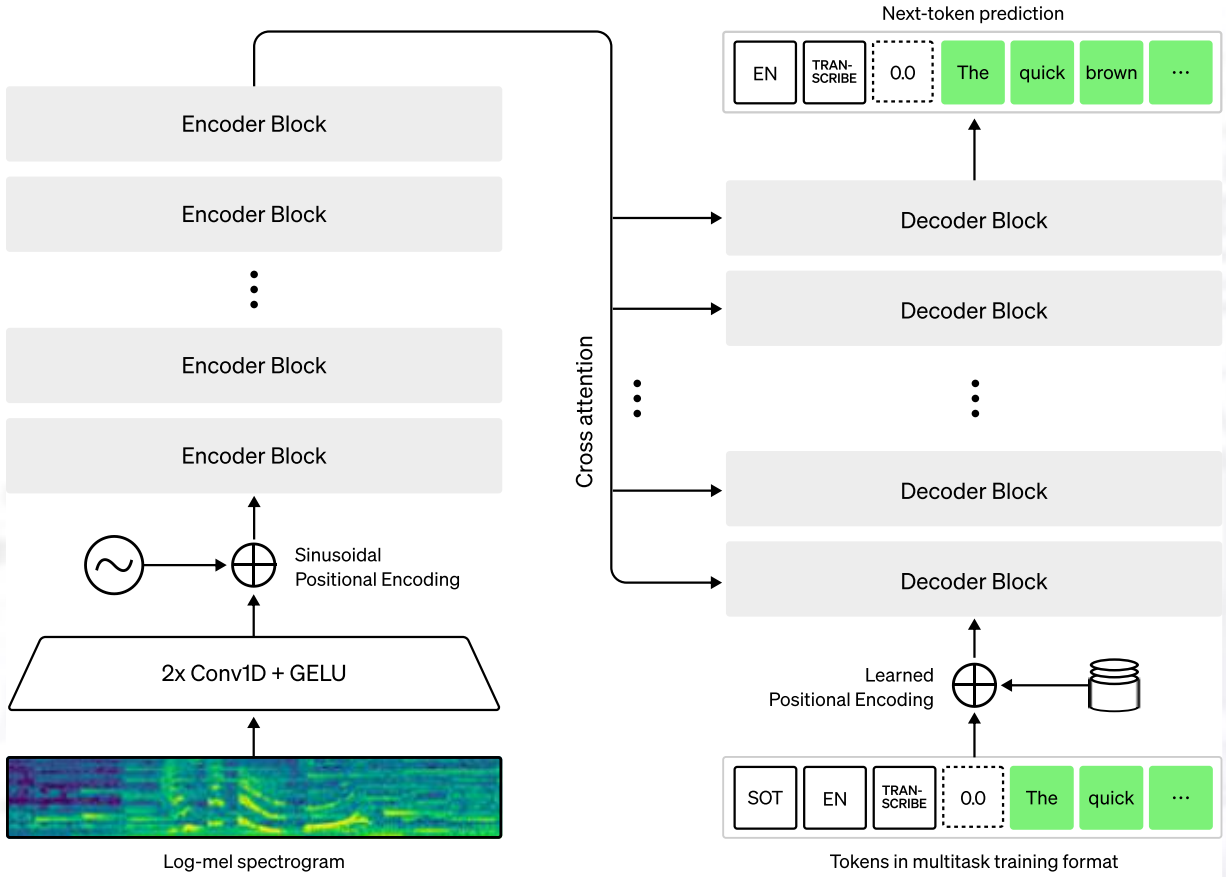
Identificación de Idioma

- **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



Etapas y Procesos RAH:





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:

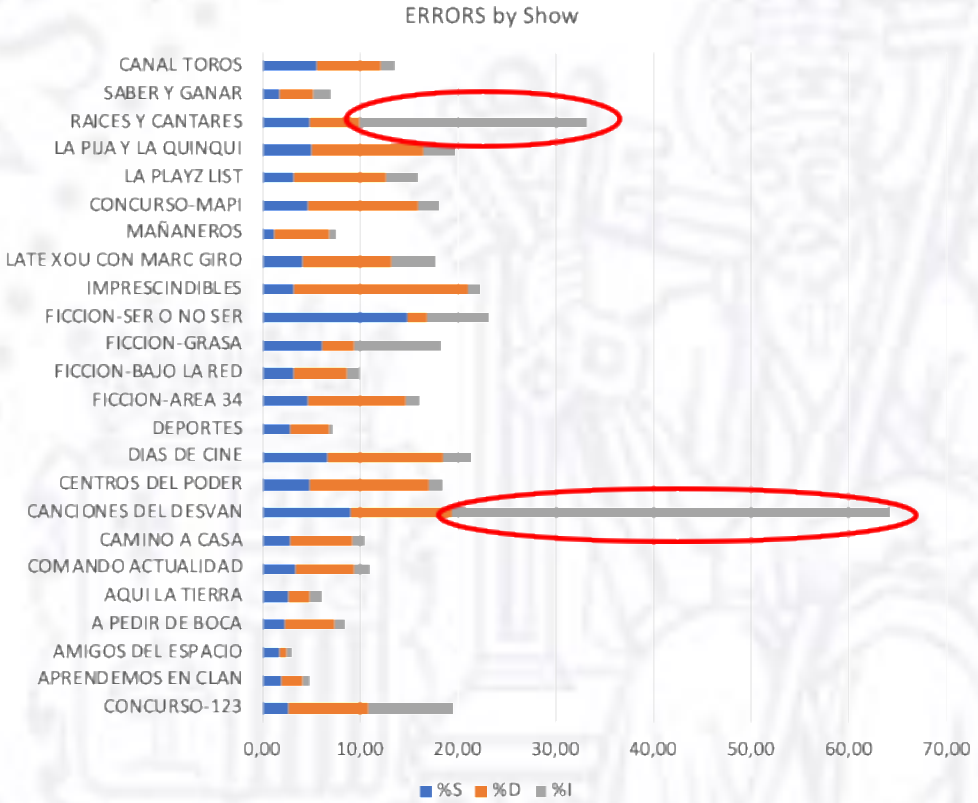
REF: a las tres y siete minutos de mañana
HYP: a las tres diecisiete minutos de la mañana

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

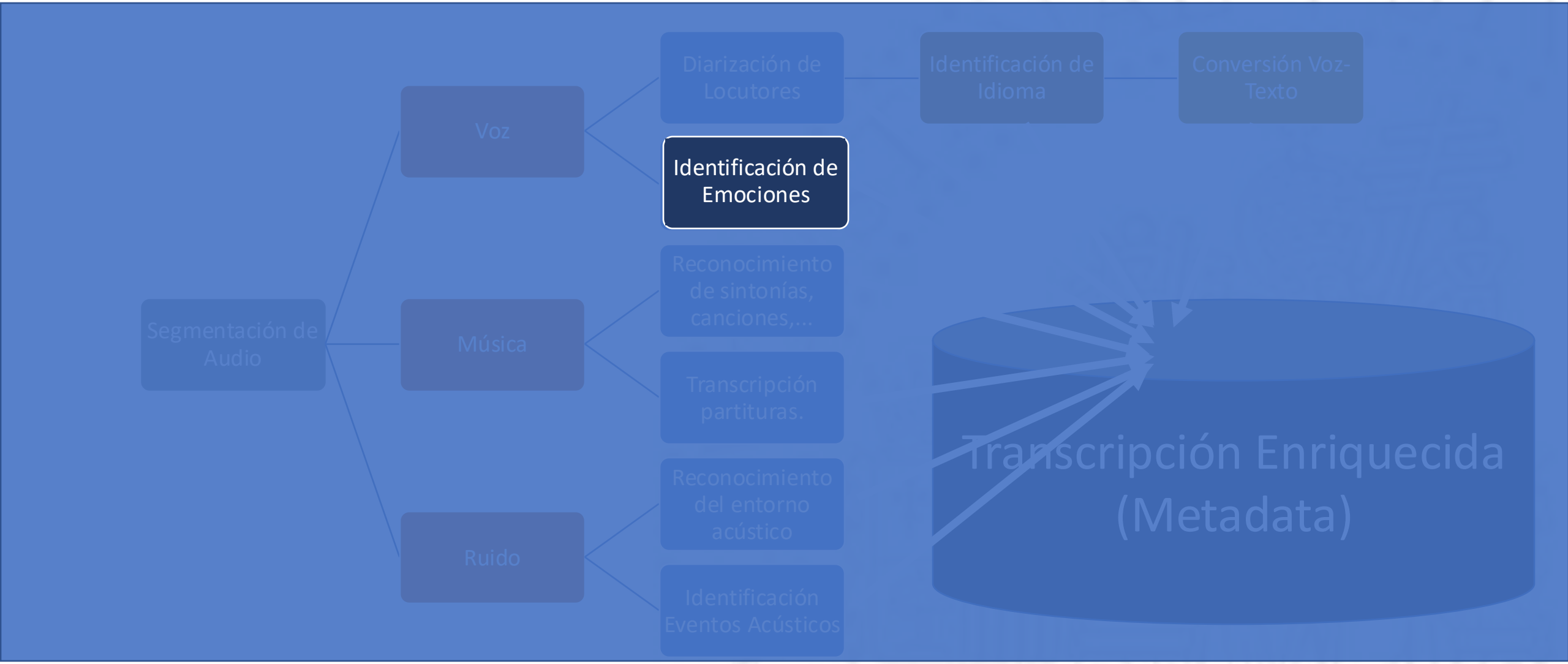


WER Baseline WhisperX large-v3: 2024: 12,10 %

Diarización e Identificación de hablantes + Reconocimiento Automático del Habla:

-  **D. Trump:** Contenido de la intervención 1 … <Tcomienzo1> <Tfin1>
-  **F. Merz:** Contenido de la intervención 2 … <Tcomienzo2> <Tfin2>
-  **K. Starmer:** Contenido de la intervención 3 … <Tcomienzo3> <Tfin3>
-  **D. Trump:** Contenido de la intervención 4 …. <Tcomienzo4> <Tfin4>
-  **F. Merz:** Contenido de la intervención 5 … <Tcomienzo5> <Tfin5>

Tecnologías

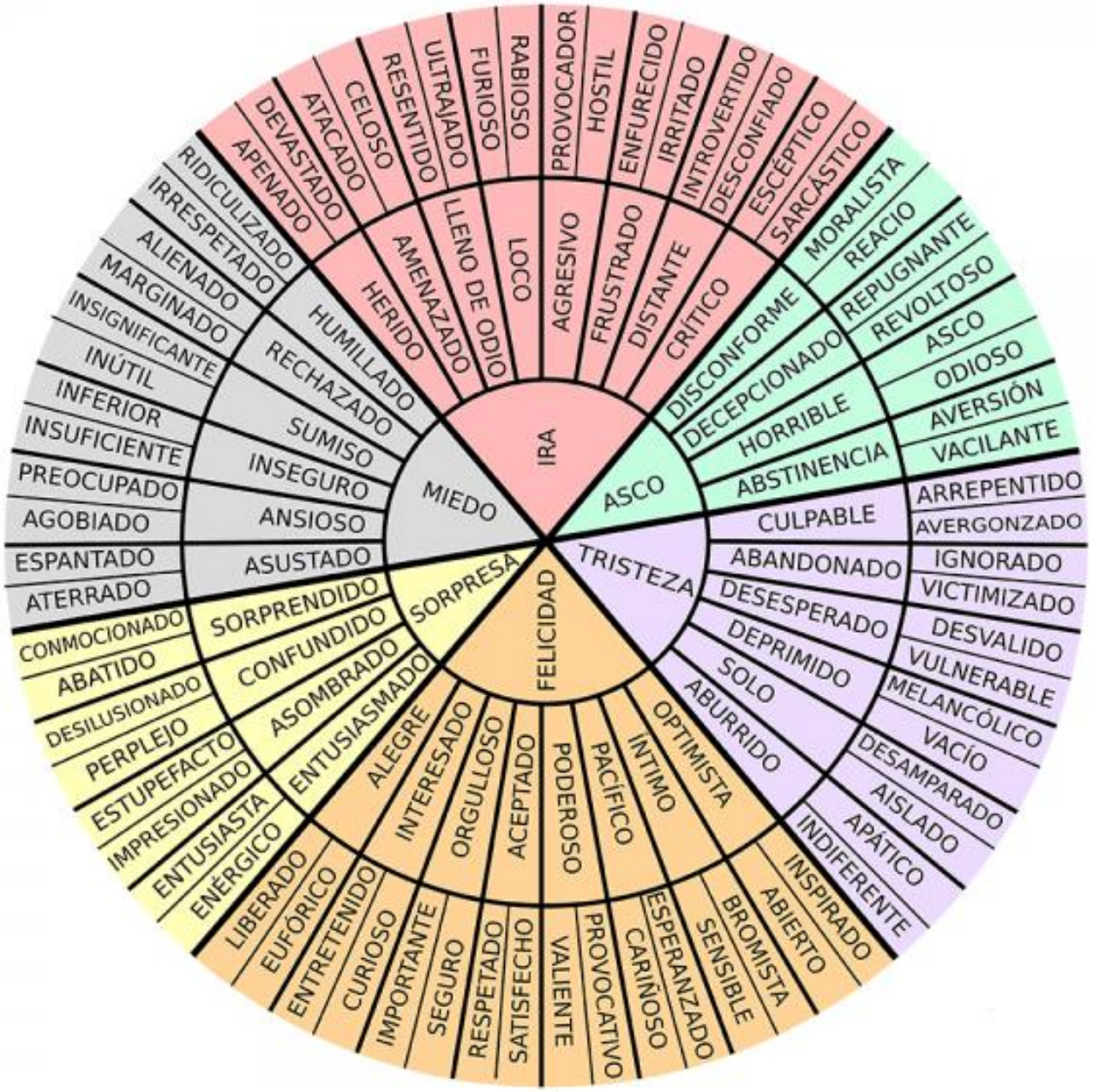


Identificación de Emociones

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

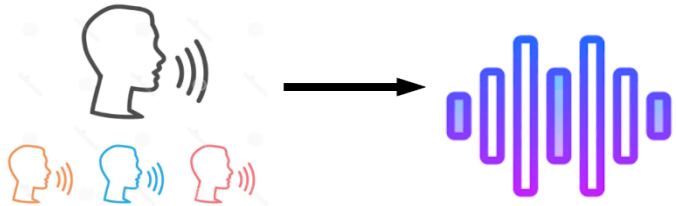
El problema de los datos

- Cantidad limitada de audio
- Emociones simuladas
- Pocos locutores
- Etiquetado subjetivo

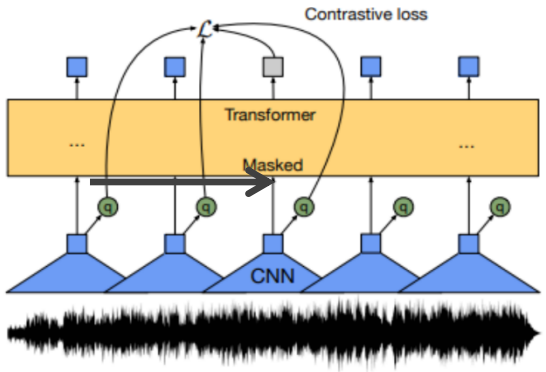


Identificación de Emociones

SEÑAL DE VOZ



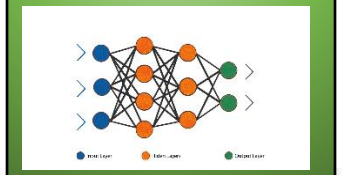
Feature Extraction Using Wav2Vec2



Classify Layer

Feature Vectors

CLASIFICACIÓN



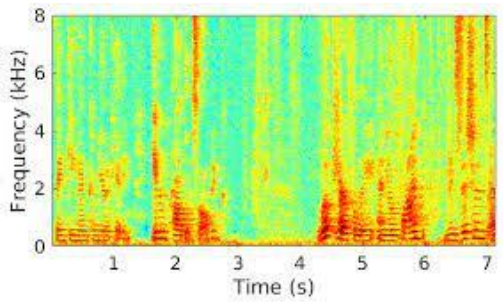
Emotion Recognition

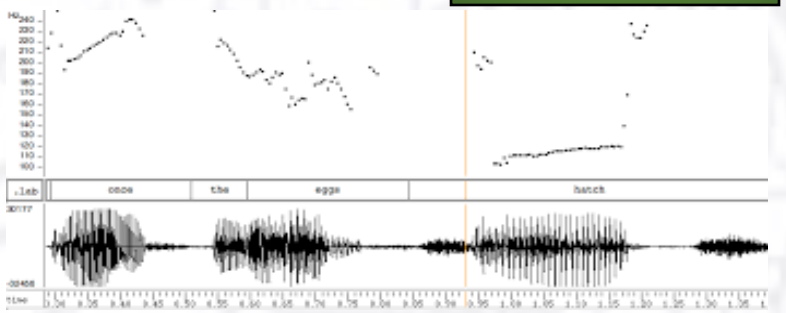
- enfado
- alegría
- tristeza
- neutro
- aburrimiento
- ansiedad

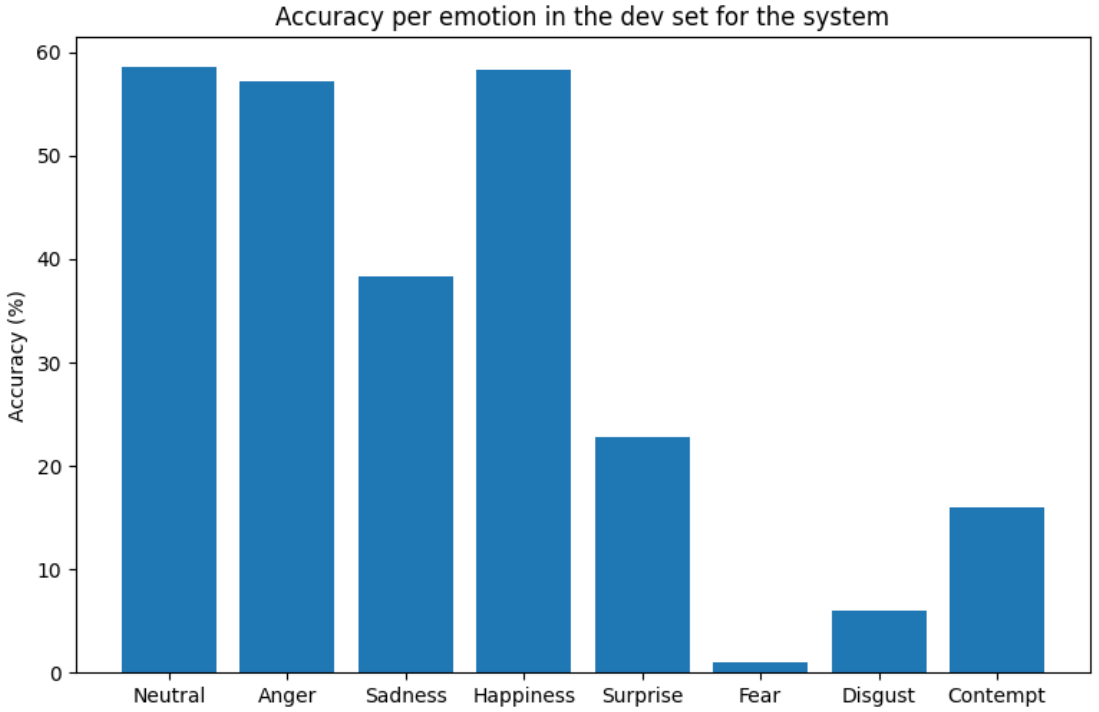
Características:

Espectrales
Prosódicas
Paralingüísticas

...







- Diferencias grandes en cuanto a precisión entre emociones
- La diferencia en precisión es consecuente con los estudios realizados en humanos y el consenso del conjunto de datos.