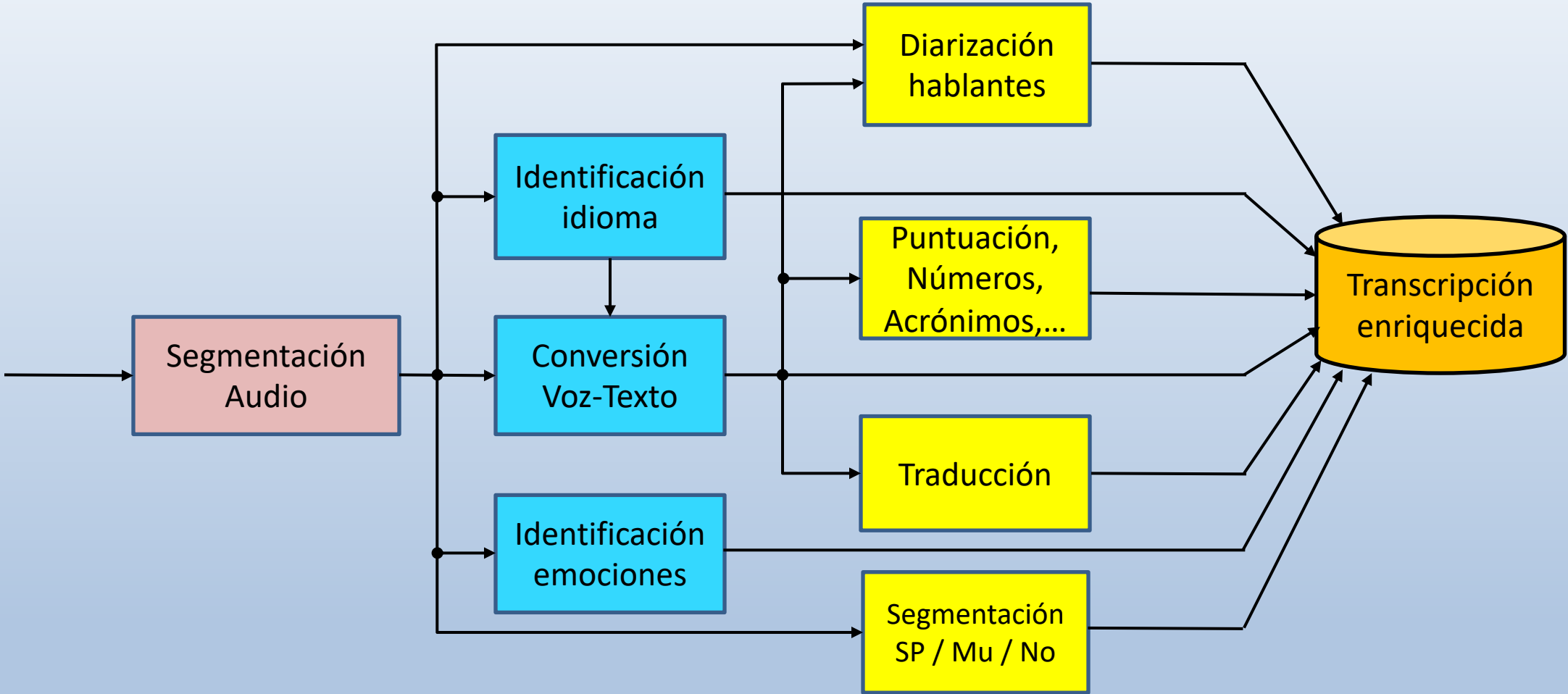




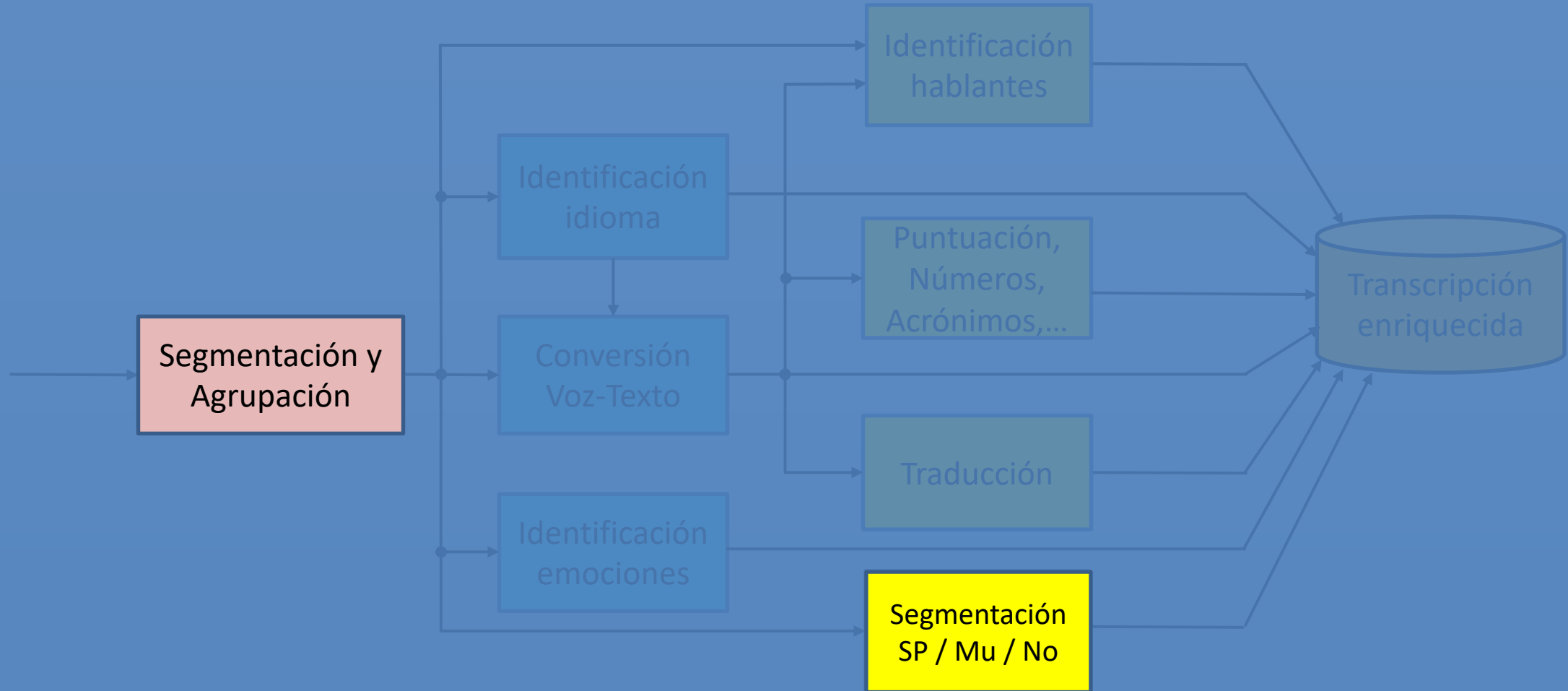
Universidad
Zaragoza

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

Ejemplo de aplicación de Tecnologías del Habla



Ejemplo de aplicación de Tecnologías del Habla



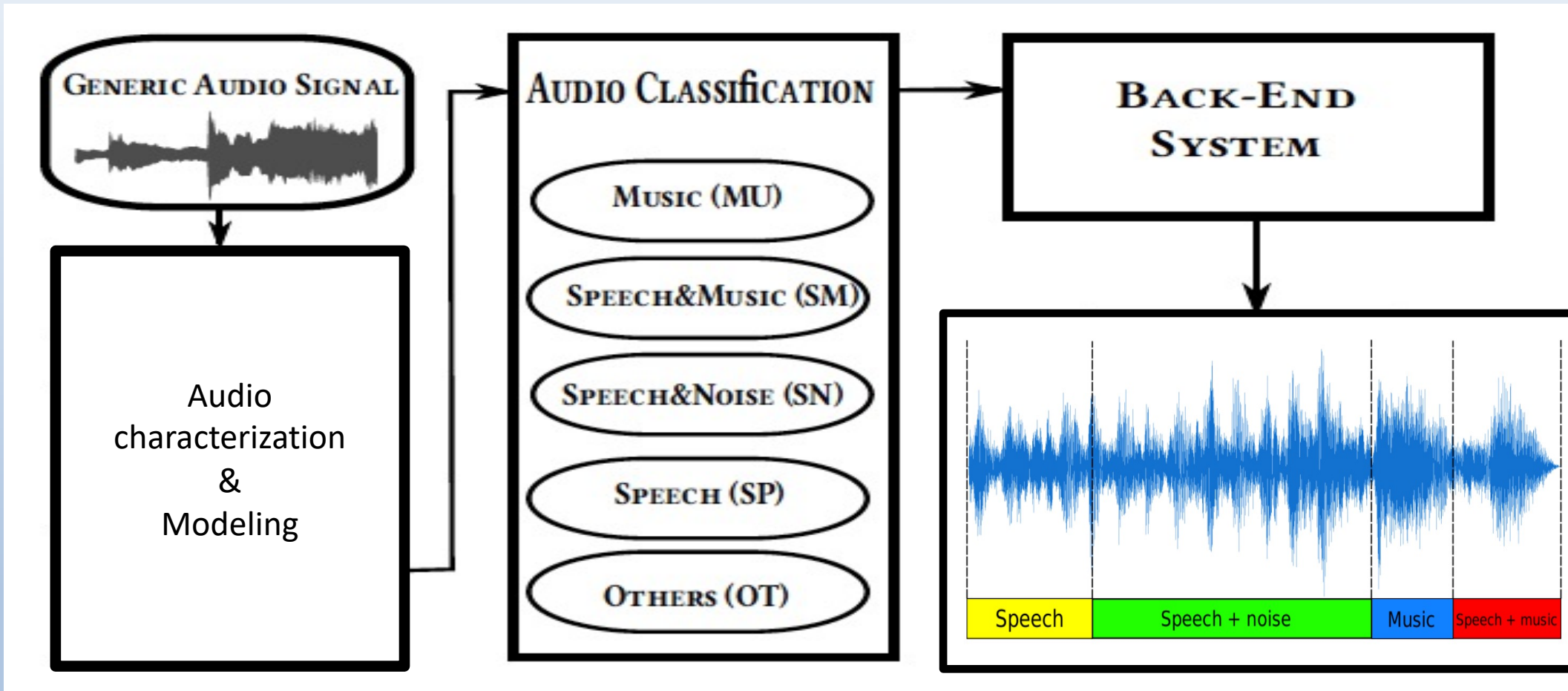
Segmentación de Audio

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

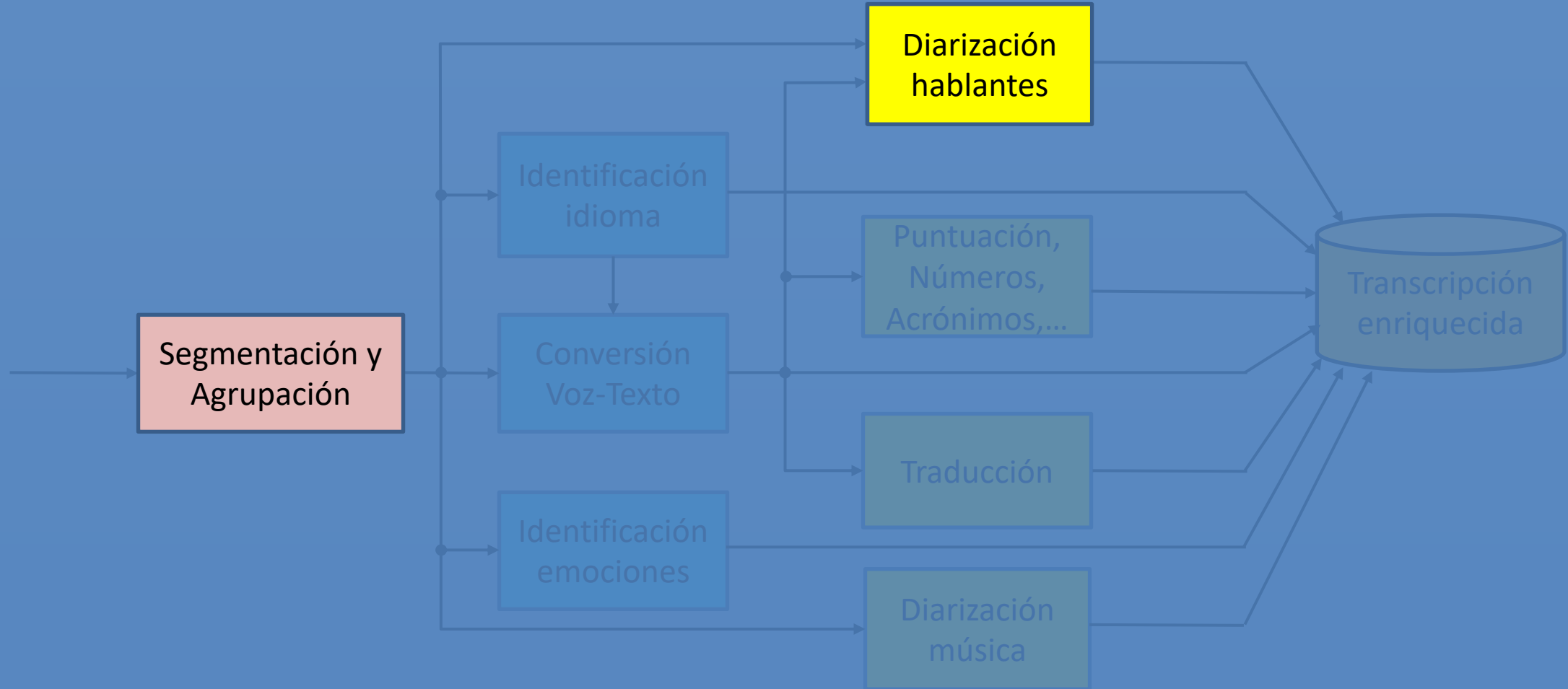
Segmentación de Audio

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

Segmentación de Audio



Ejemplo de aplicación de Tecnologías del Habla



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.*

Segmentación y Agrupación de Hablantes



Segmentación y Agrupación de Hablantes

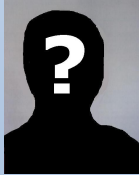
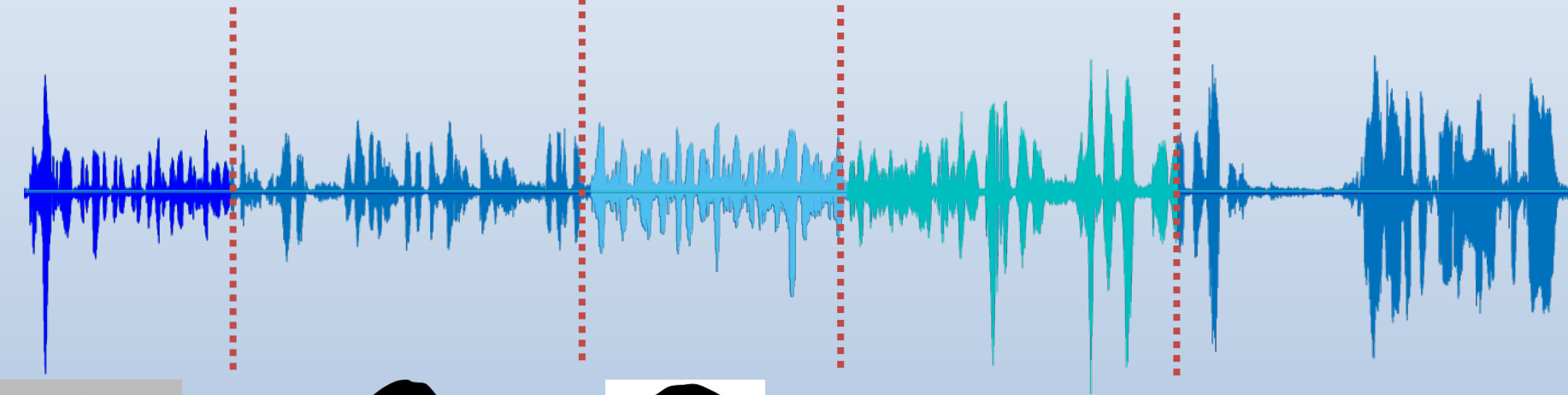
Segm 1

Segm 2

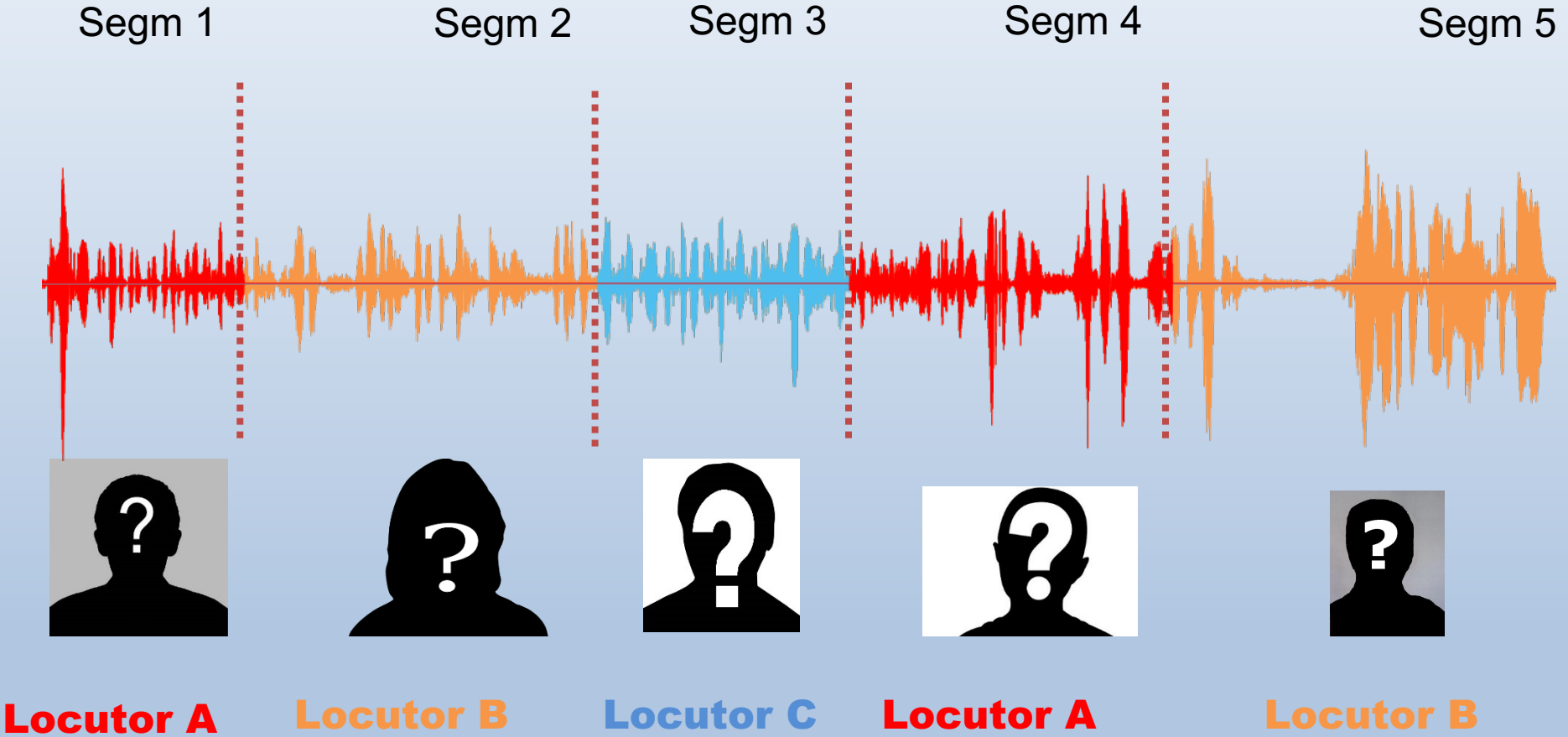
Segm 3

Segm 4

Segm 5



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:

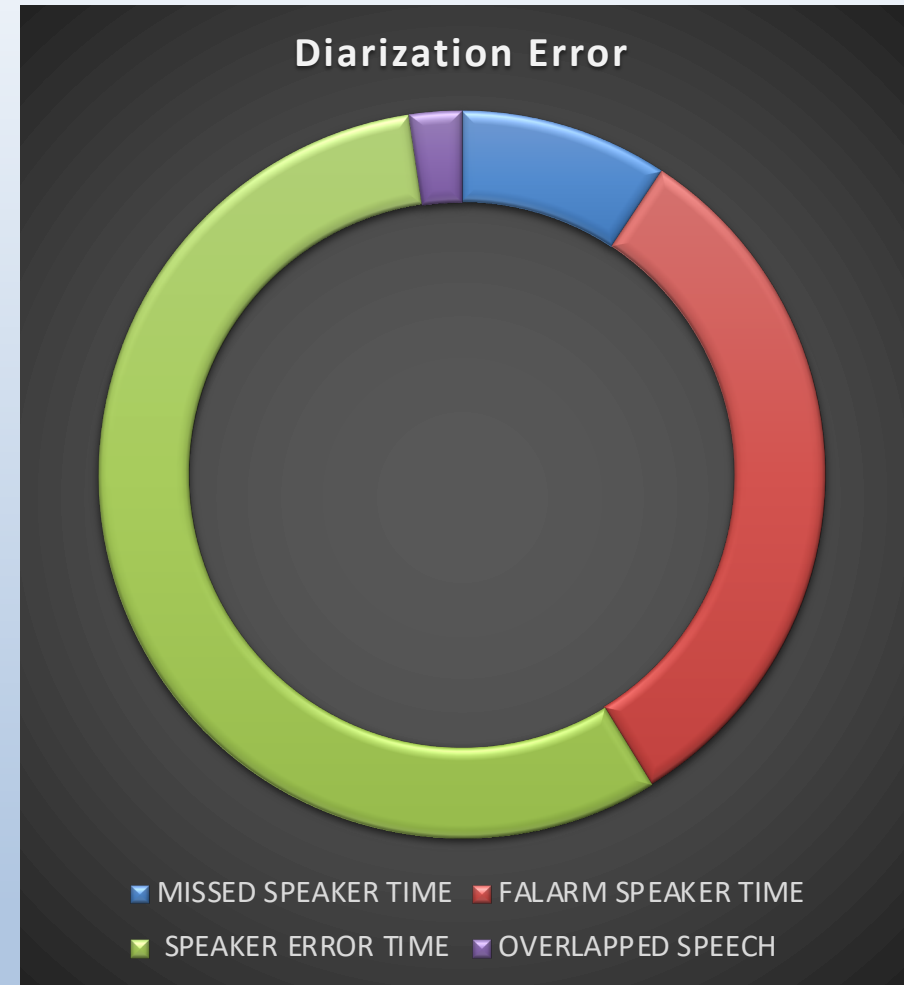
- No hay voz pero se ha detectado erróneamente.

- Error de Locutor:

- Se ha confundido un locutor con otro.

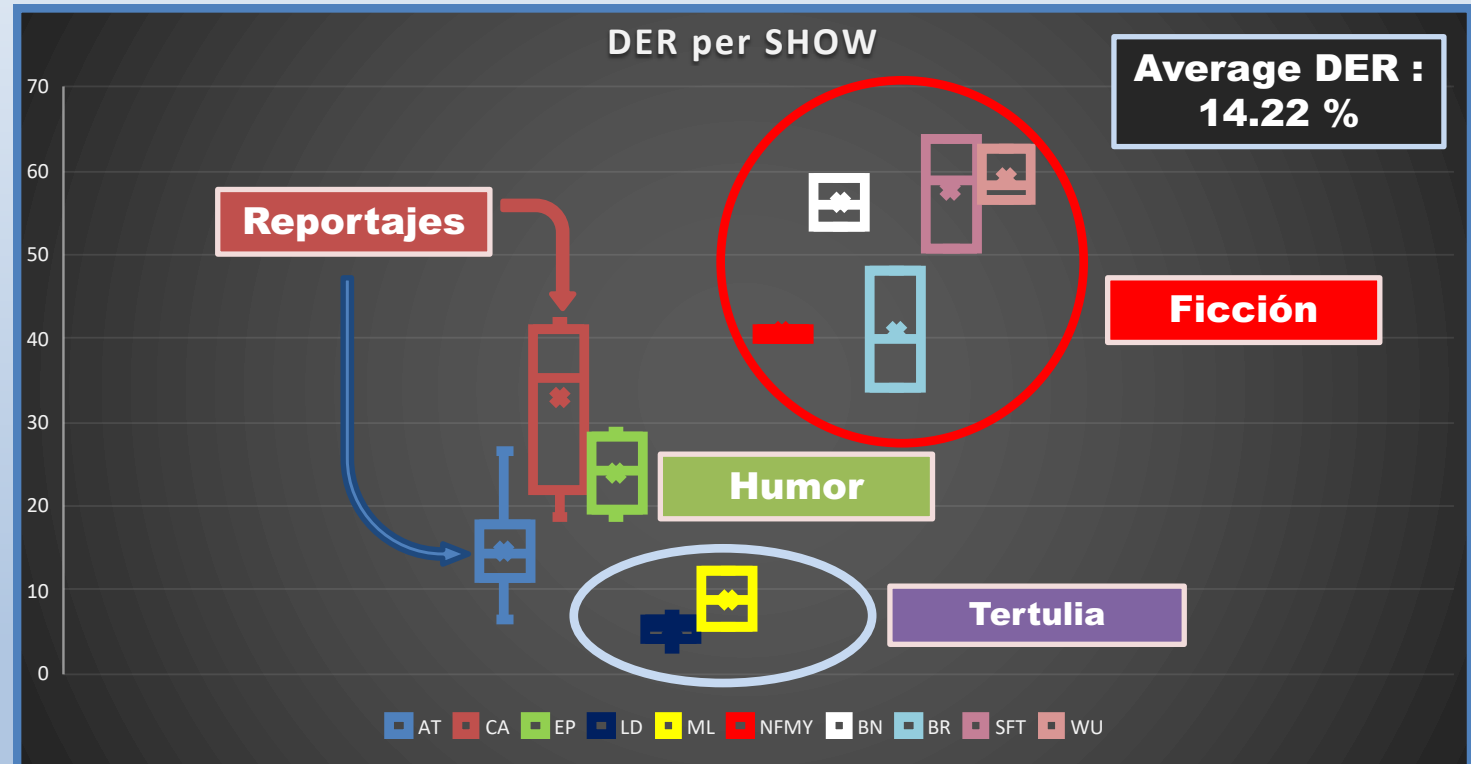
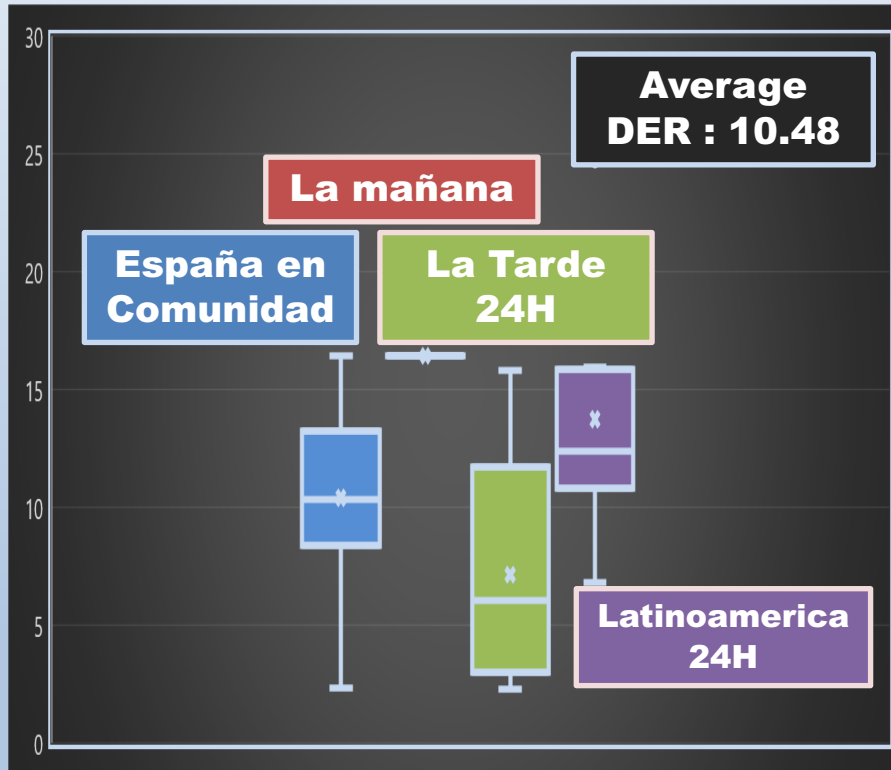
- Error por Solape:

- Dos locutores hablan a la vez, pero solo se ha identificado a uno.

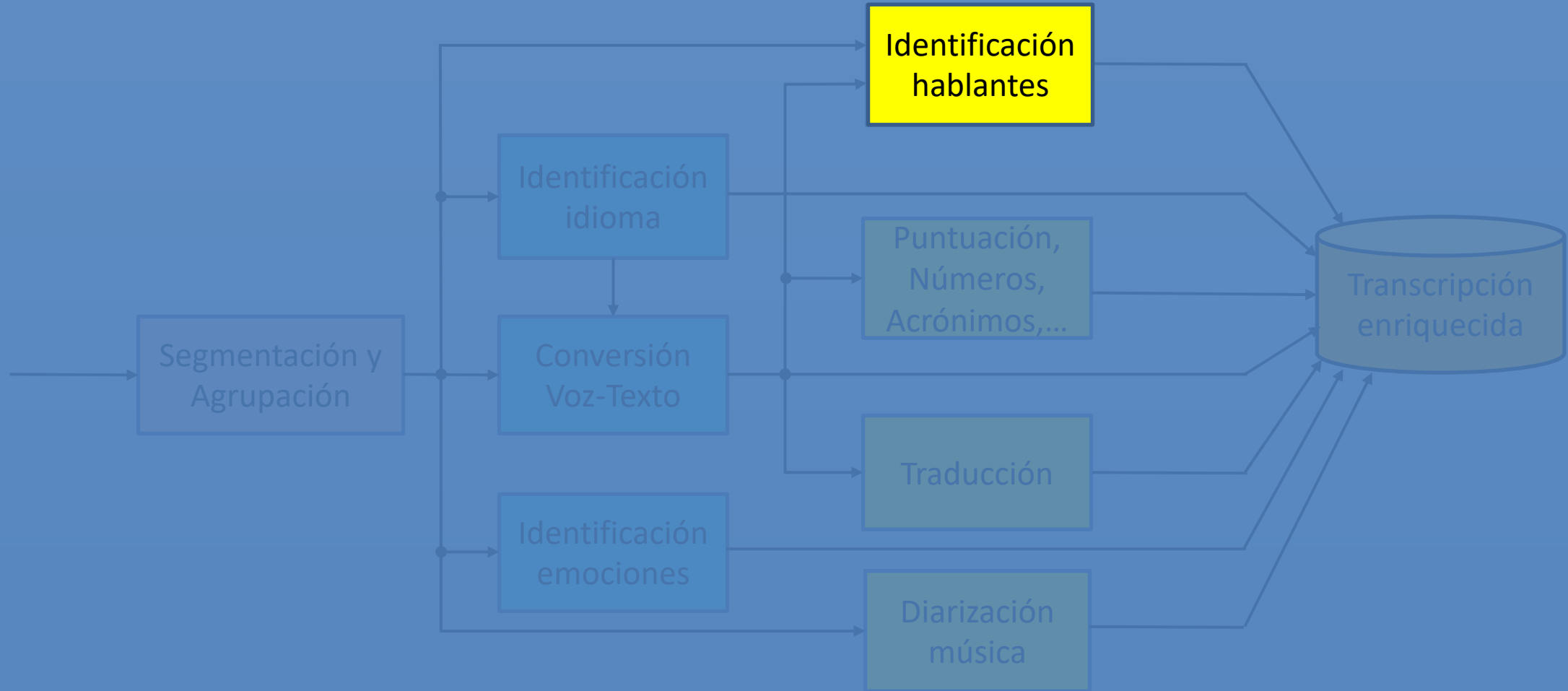


Prestaciones: Albayzin 2018 y 2020

- Broadcast:



Ejemplo de aplicación de Tecnologías del Habla

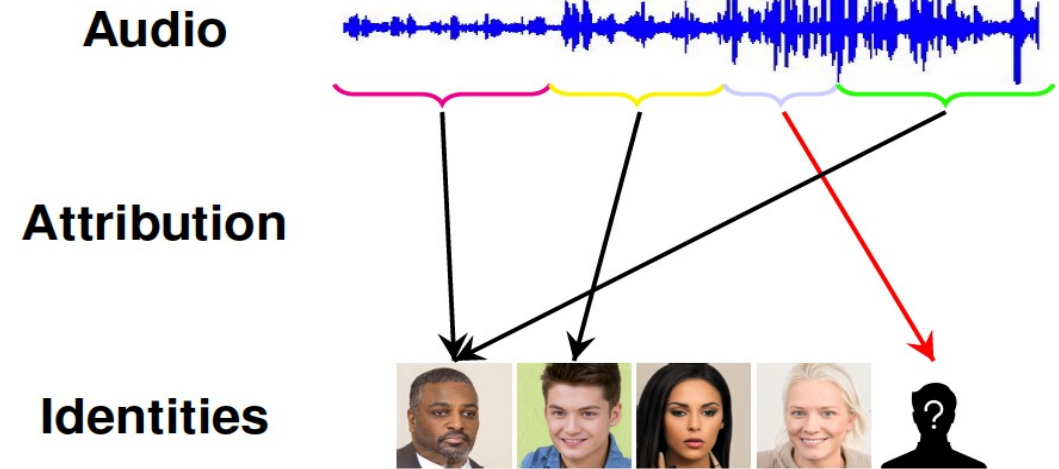
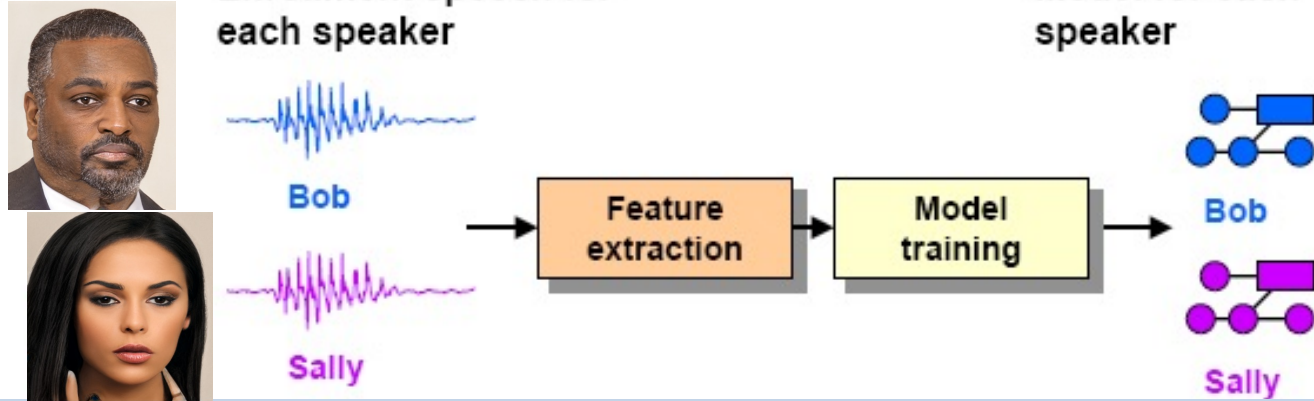


Identificación de Hablantes

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

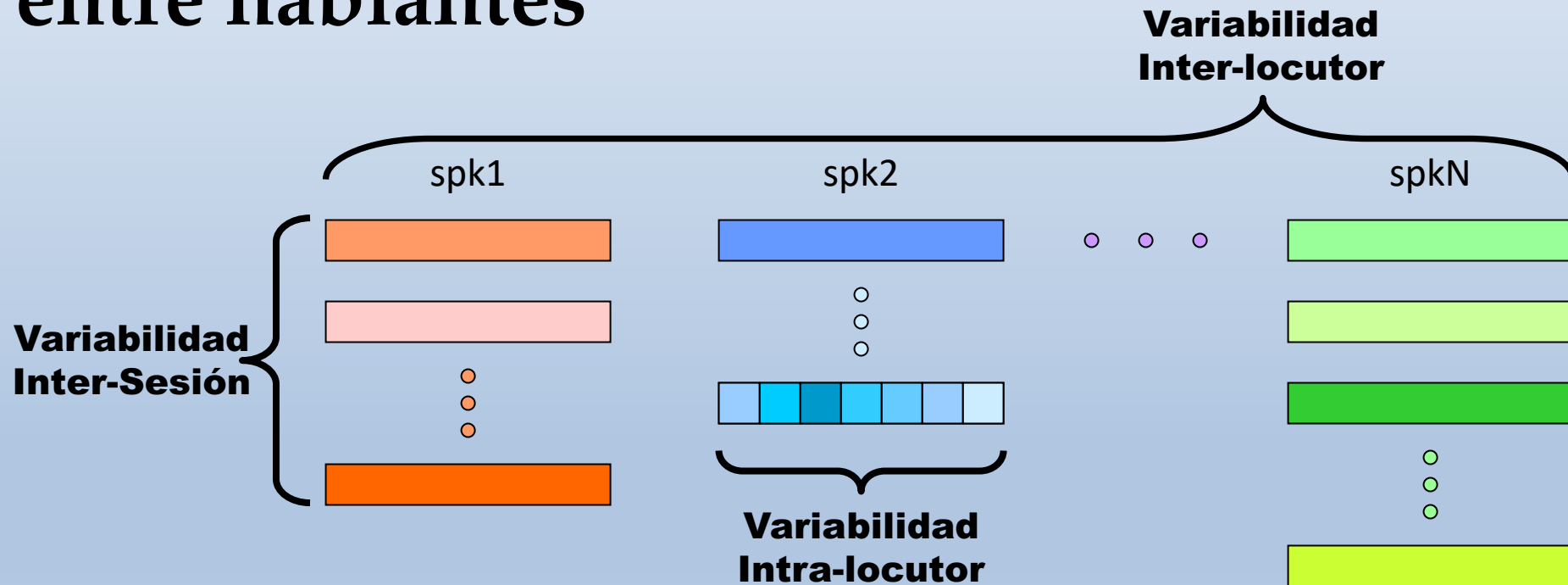
Identificación de Hablantes

Enrollment Phase



Retos en la Identificación de hablantes

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



Junto con la Identificación del Hablante:

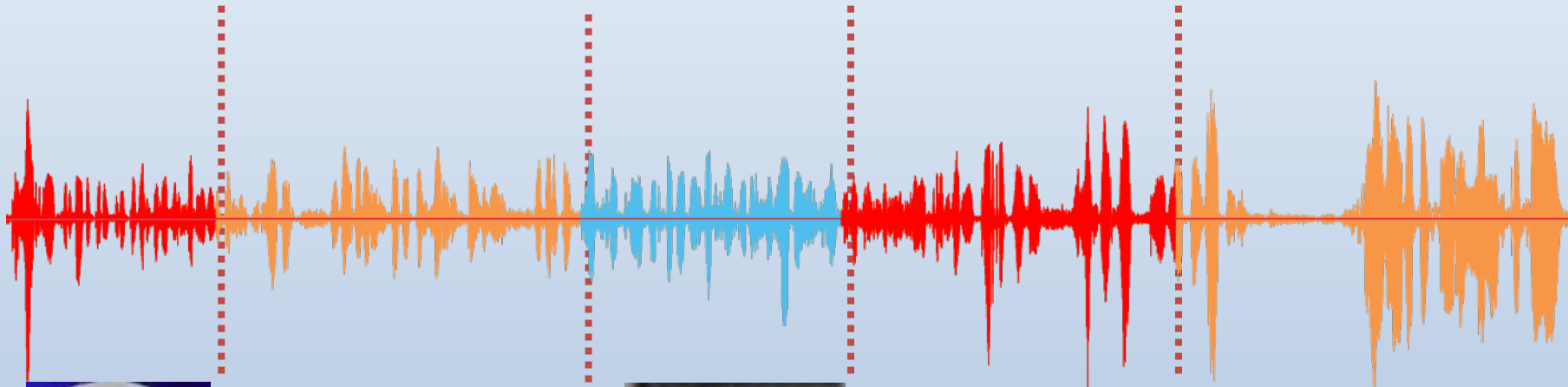
Segm 1

Segm 2

Segm 3

Segm 4

Segm 5



Joe Biden A



Donald Trump B



Boris Johnson



Joe Biden A



Donald Trump B

Prestaciones: Albayzin 2020

- Iberspeech-RTVE-Challenge :*

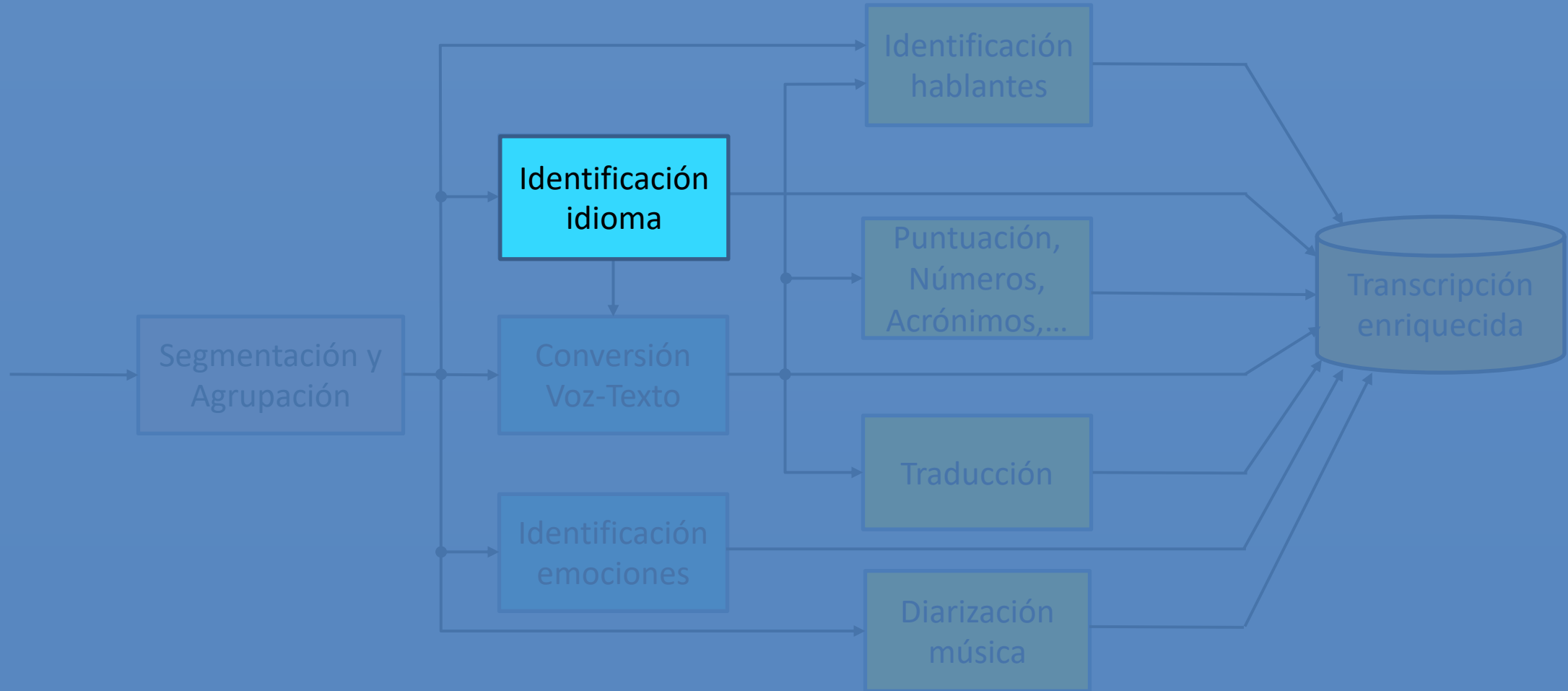


	AER
BIOMETRIC VOX	65.09 %
VIVOLAB	72.63 %

	MISS	FA	SPK ERR
BIOMETRIC VOX	47.0 %	9.2 %	8.9 %
VIVOLAB	5.1 %	53.3 %	14.2 %

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

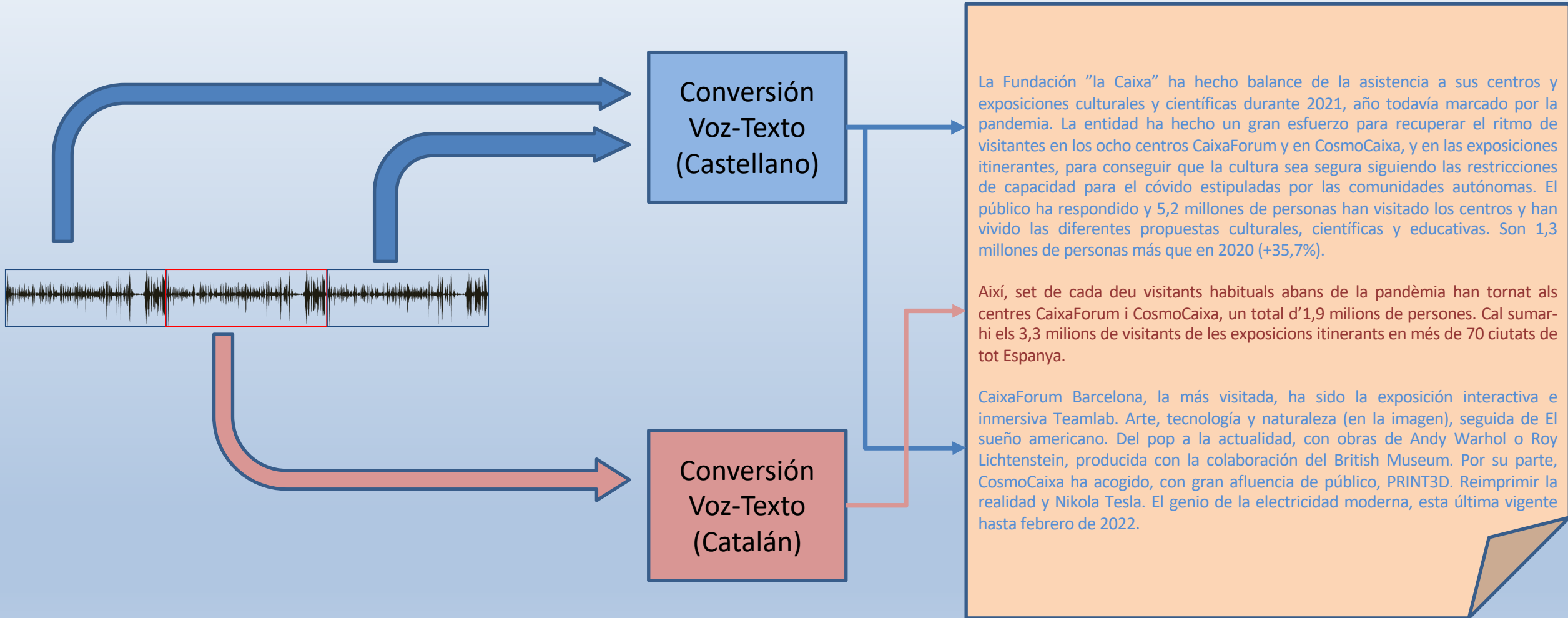
Ejemplo de aplicación de Tecnologías del Habla



Identificación de Idioma

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

Identificación de Idioma



Identificación de Idioma

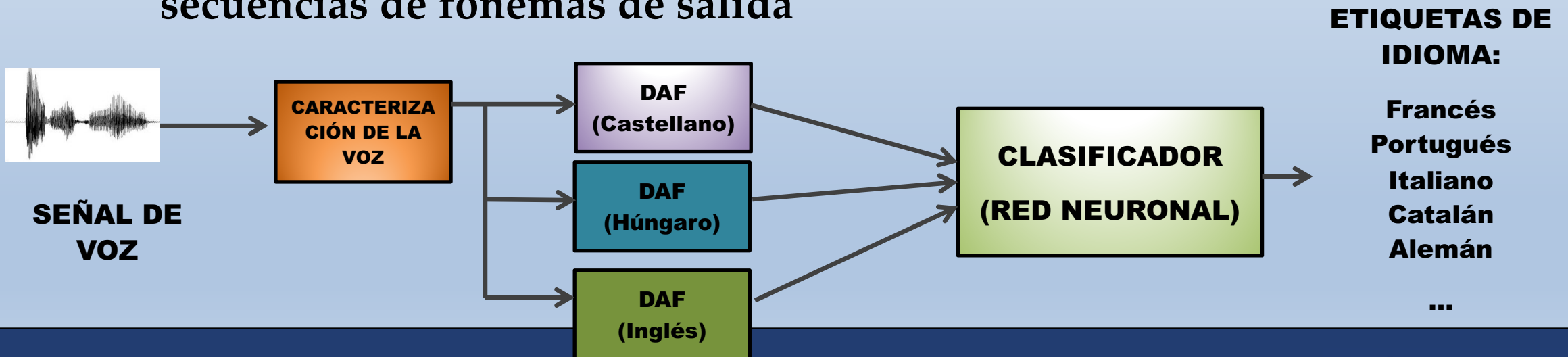
- **Tecnologías:**

- Acústicos

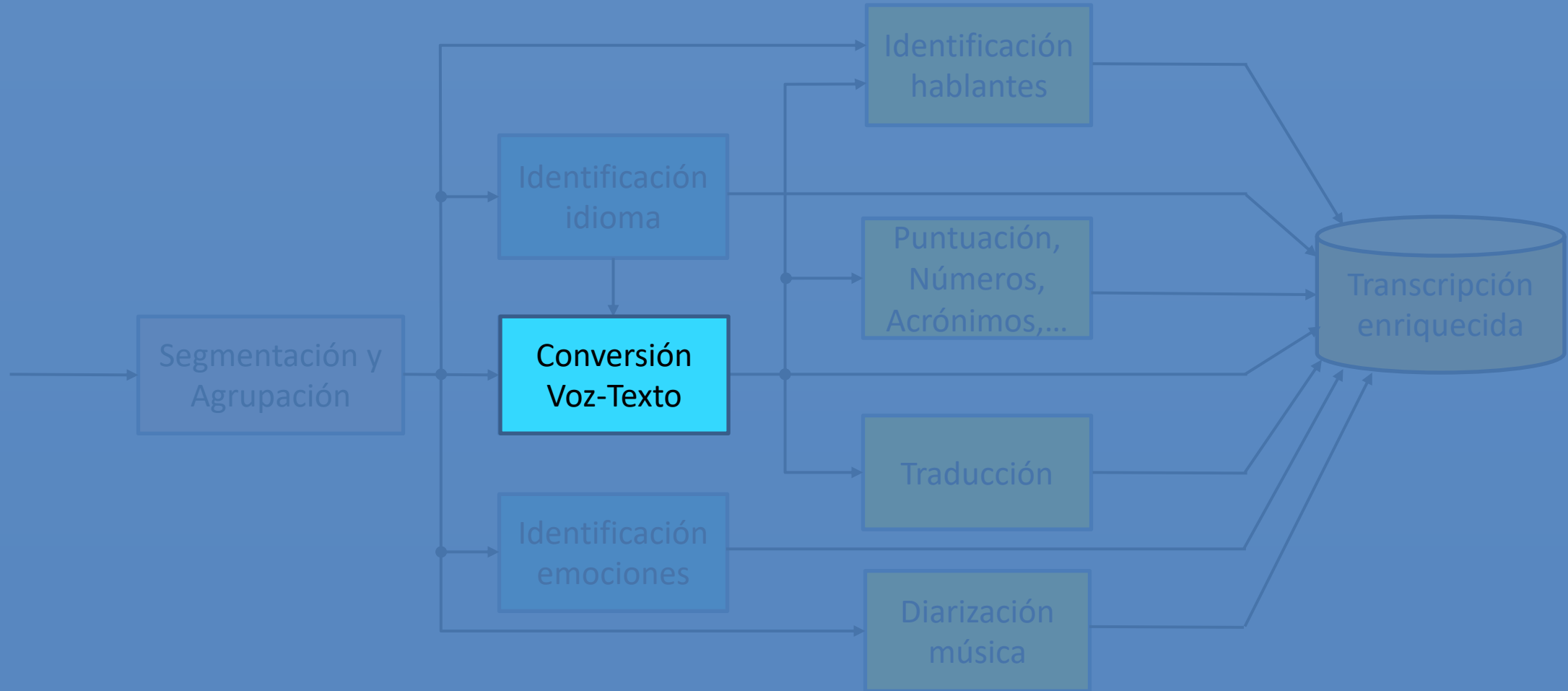
- Tratan de buscar patrones discriminativos directamente sobre la señal de voz

- Fonotácticos (Lingüísticos)

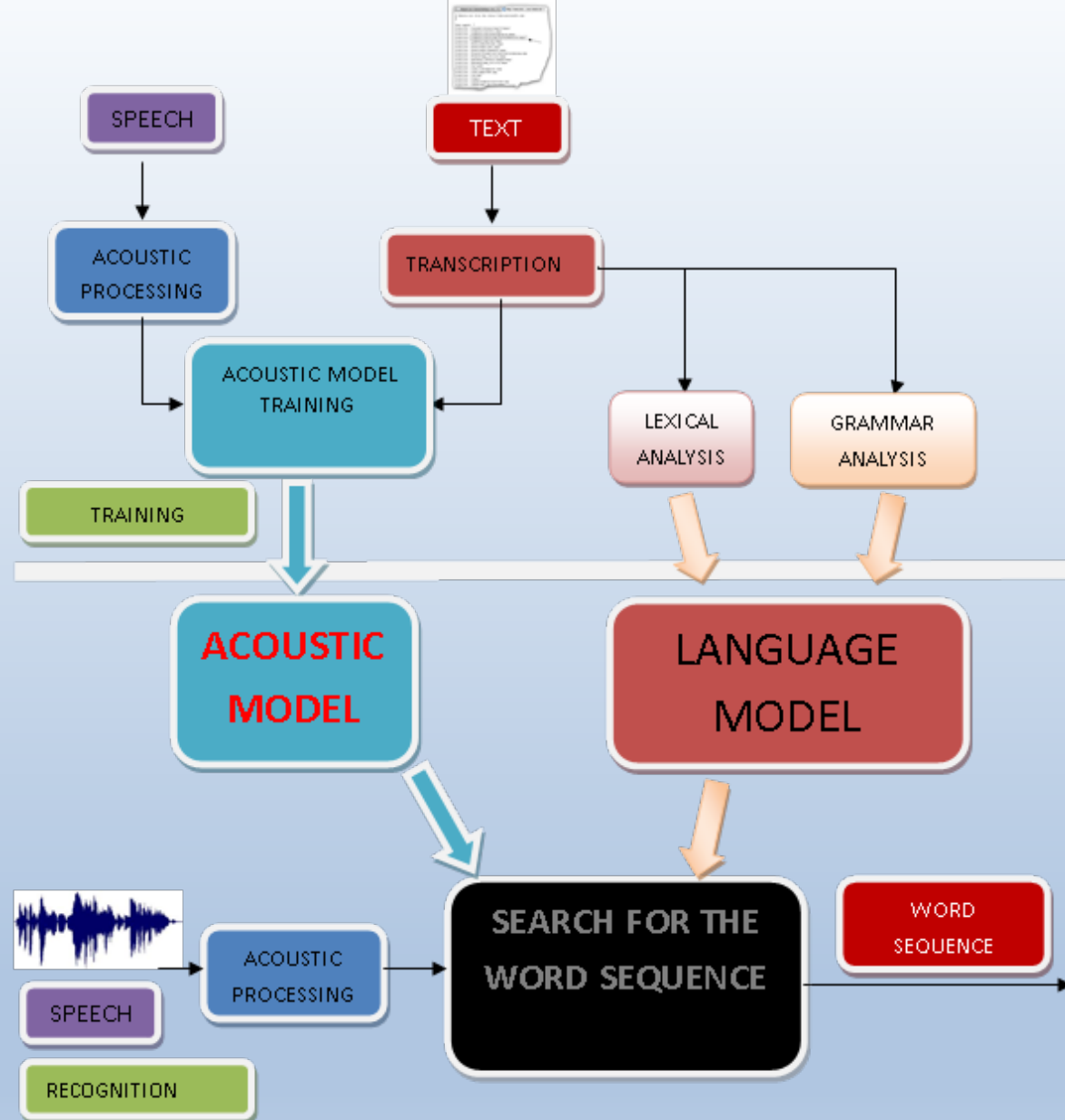
- Primero procesan la señal de entrada con un (o varios) reconocedor fonético (en varios idiomas) después buscan patrones discriminativos en las secuencias de fonemas de salida



Ejemplo de aplicación de Tecnologías del Habla



Etapas y Procesos Reconocimiento Automático del Habla:



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 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 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:

Substituciones (S), Borrados (B), Inserciones (I)

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

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

EL MODELO LÉXICO Y MODELO DE LENGUAJE

Modelo léxico

Relación entre las unidades lingüísticas y las palabras

apagar a p a G 'a r

bautizo b a w t 'i T o

candado k a n d 'a D o

Modelo de lenguaje

Relación entre las palabras y las frases

Gramáticas formales (regulares, contexto libre,...)






Modelos probabilísticos (N-gramas, redes neuronales)

$P(\text{la, alta, comisionada, de, las, naciones, unidas}) = P(\text{la}) \times P(\text{alta|la}) \times P(\text{comisionada|la,alta}) \times$
 $P(\text{de|la,alta,comisionada}) \times P(\text{las|la,alta,comisionada,de}) \times P(\text{naciones|la,alta,comisionada,de,las})$
 $\times P(\text{unidas|la,alta,comisionada,de,las,naciones})$

¿Cómo se aprende?

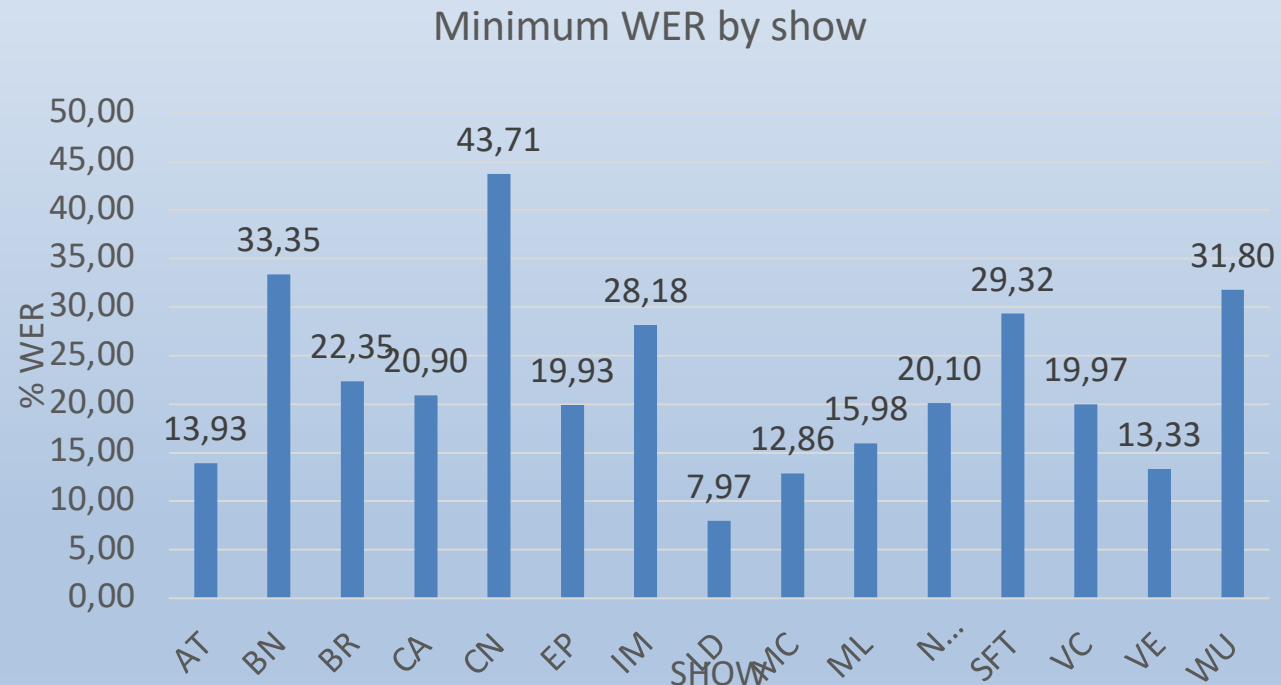
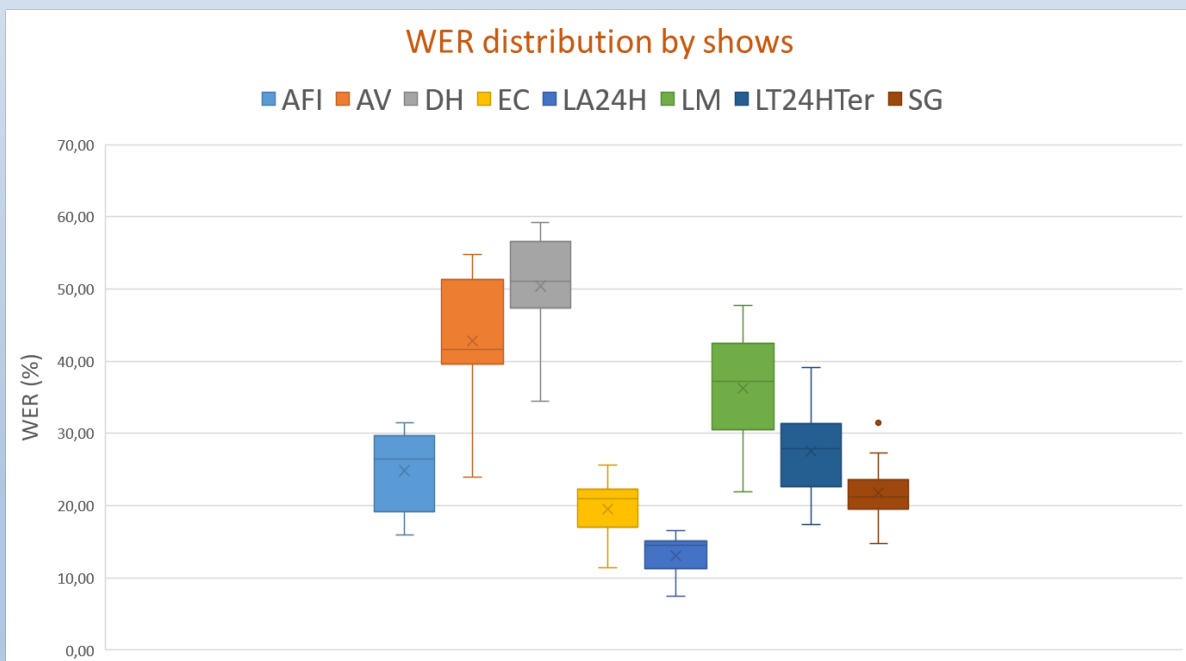
Millones de textos

Añadiendo Reconocimiento del Habla:

- ◉  **J. Biden:** Contenido de la intervención 1 ... <Tcomienzo1> <Tfin1>
- ◉  **O. Scholz :** Contenido de la intervención 2 ... <Tcomienzo2> <Tfin2>
- ◉  **B. Johnson:** 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>

Prestaciones: Albayzin 2018 y 2020

- *Iberspeech-RTVE-Challenge* :



Ejemplos de audios de evaluación:

ESPAÑA EN COMUNIDAD



del río hay muchísima la margarita africana de las entre las dunas y bueno así muy bonita y la uña de gato la red natura dos mil a la que pertenece a la playa de fresh urfé es una red europea de espacios protegidos diseñada para asegurar la supervivencia a largo plazo de las especies y los hábitats naturales en el territorio europeo además esta playa toda la zona costera de navia y el río están incluidas en la zona especial de conservación y de protección para las aves

LA TARDE EN 24H TERTULIA



al otro lado está esperanza aguirre fuera de la política también todos sabemos porque ignacio gonzález no está no estaba en la foto en fin para cuando la responsabilidad políticas hasta donde tiene que llegar el nivel del lodazal de lo que estamos viendo para que los responsables políticos últimos den la cara y asuman la responsabilidad que les toque es verdad que estamos todavía en una fase en este caso concreto de la gürtel valenciana sólo de juicio



Universidad
Zaragoza

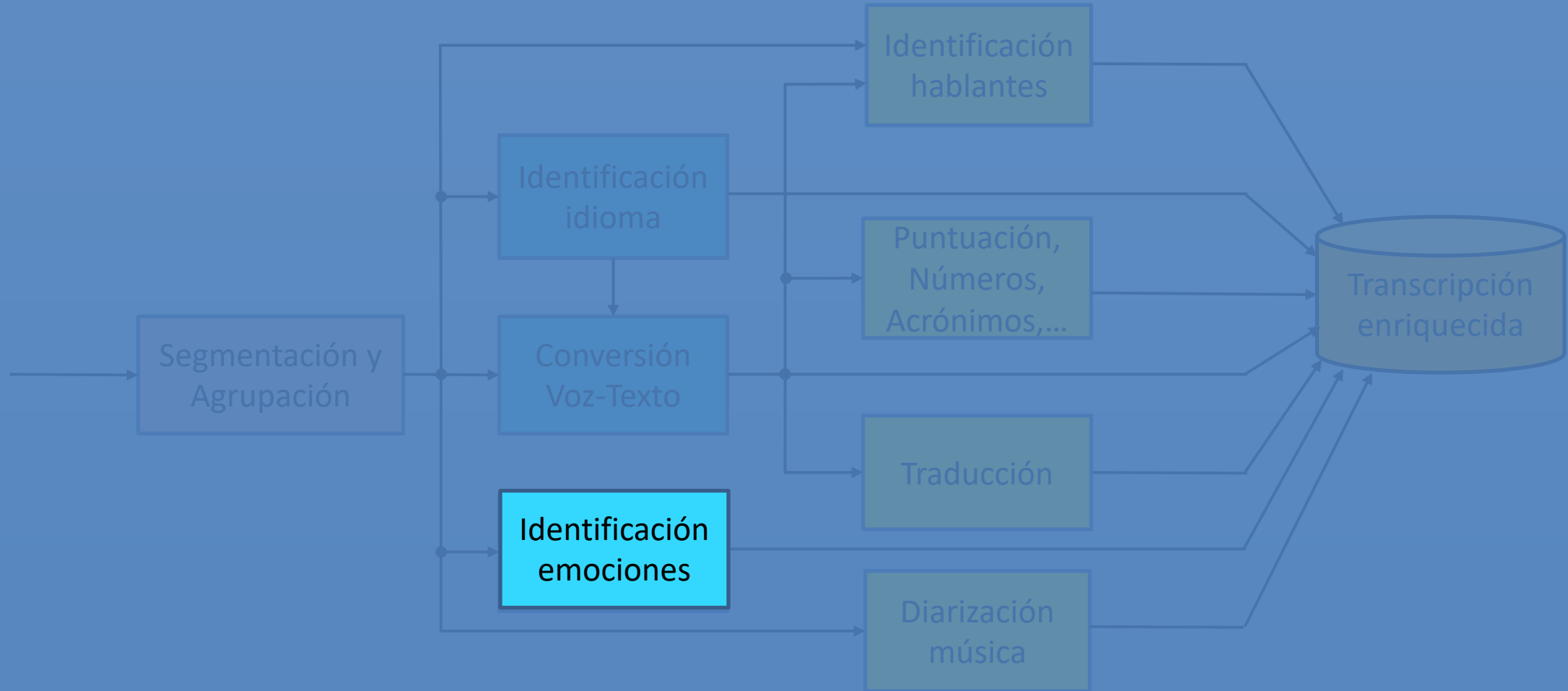


VIVOLAB



Universidad
Zaragoza

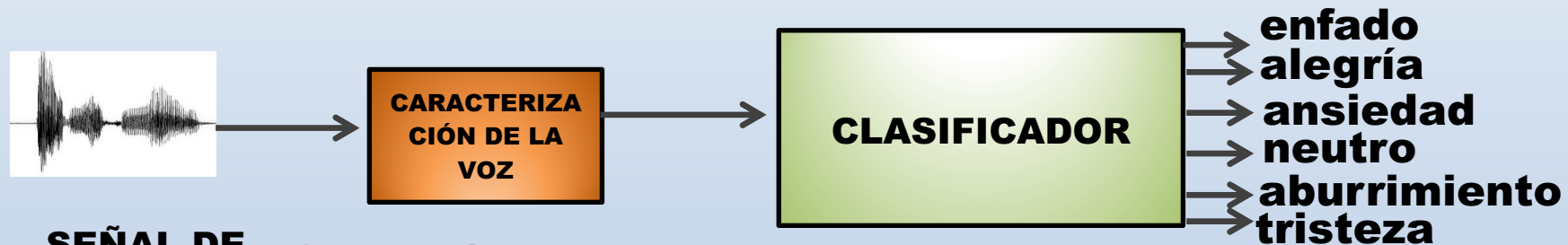
Ejemplo de aplicación de Tecnologías del Habla



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

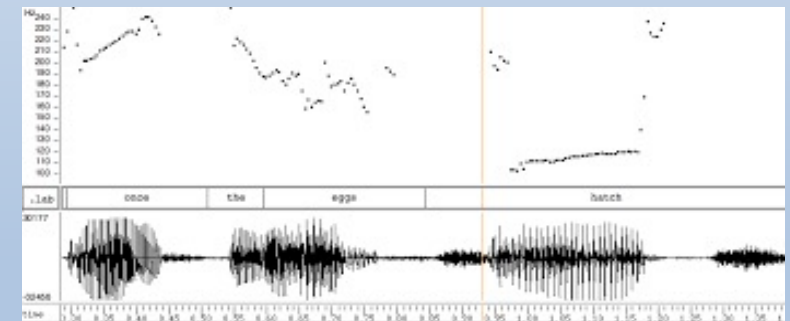
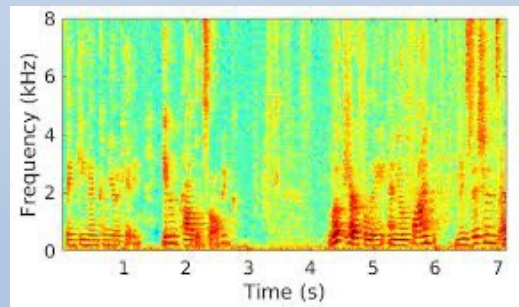
Características:

Espectrales

Prosódicas

Paralingüísticas

...





Universidad
Zaragoza