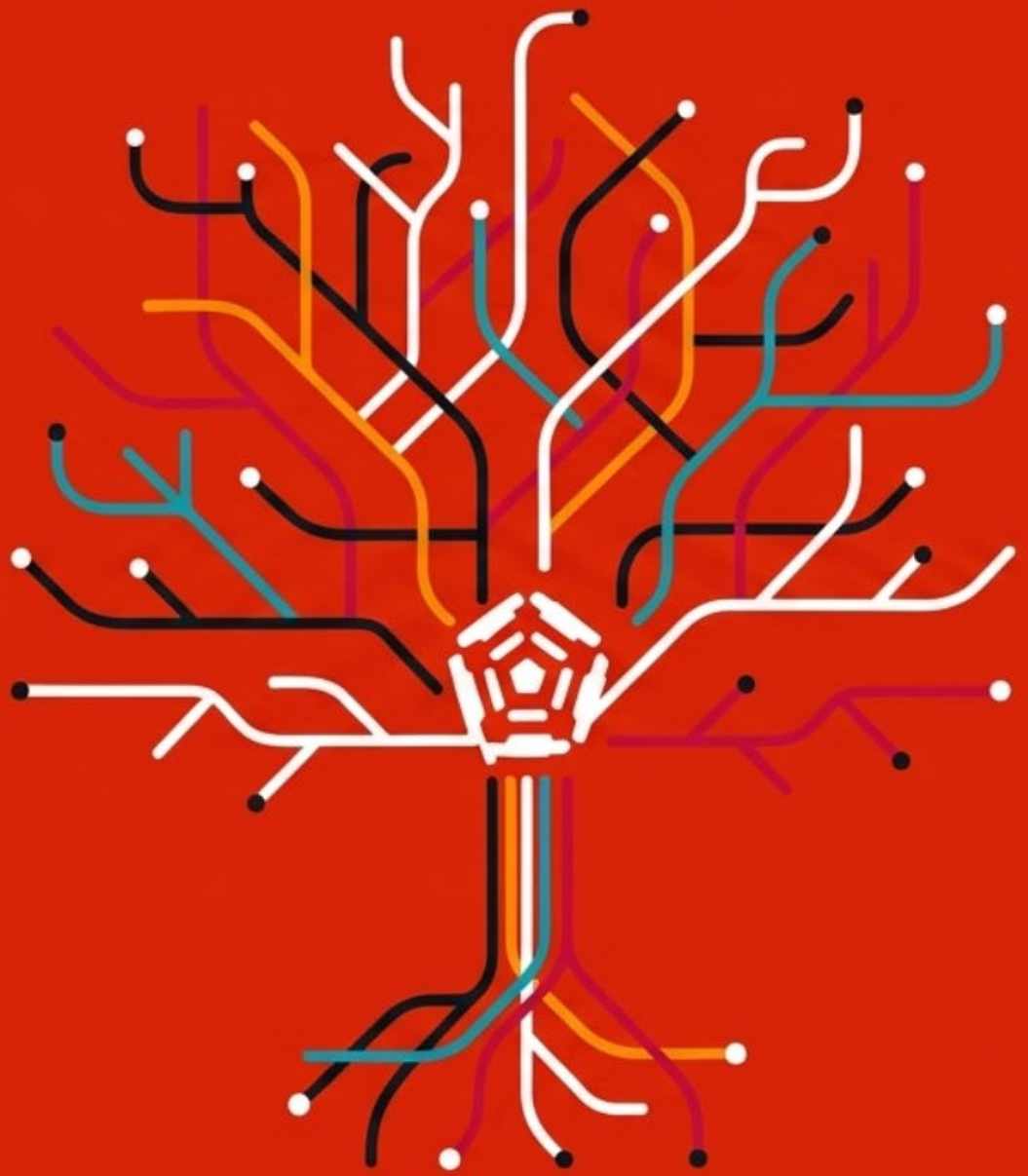


Universidad de Zaragoza | Curso 2026

¿QUÉ TIENE LA INTELIGENCIA ARTIFICIAL PARA MÍ ARCHIVO?

De la teoría a la práctica.

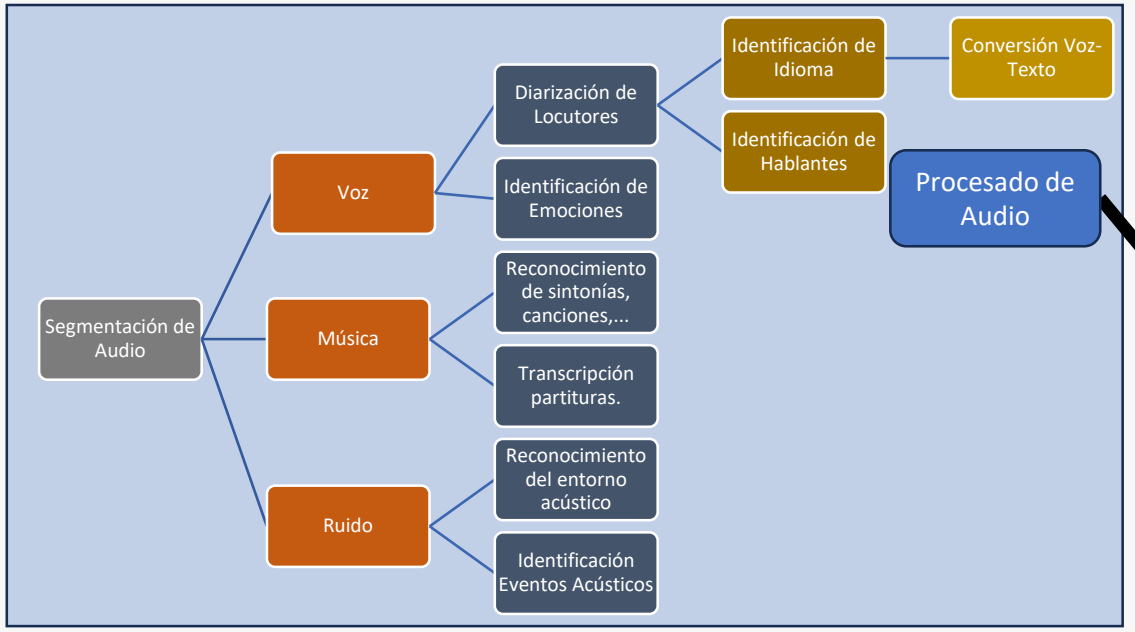
Cursos Extraordinarios
Universidad de Zaragoza 2026

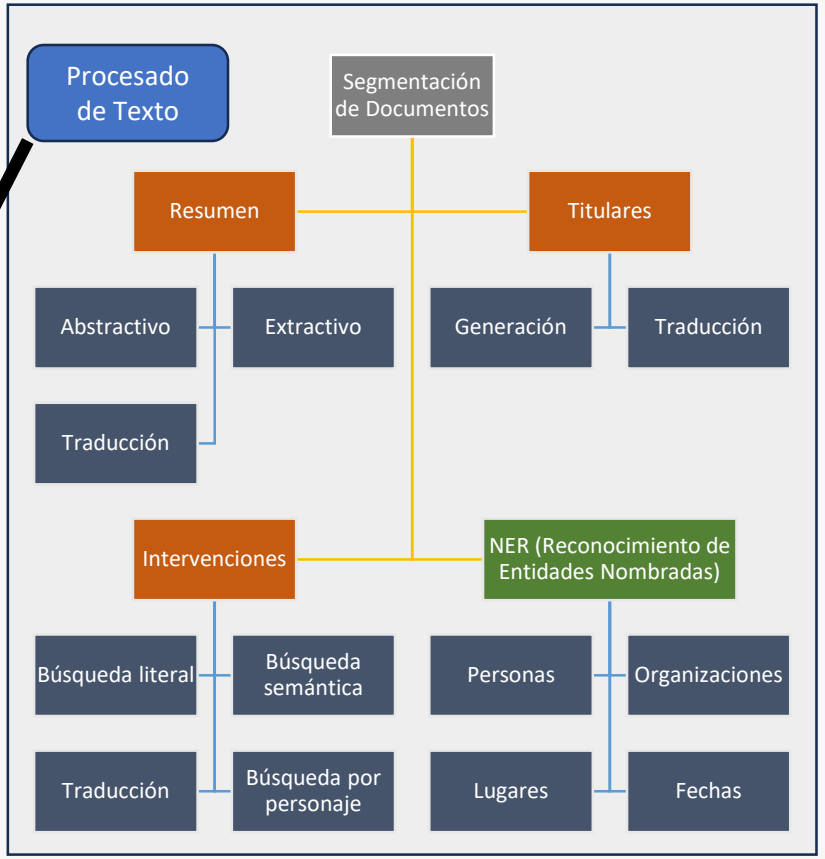
¿Qué tiene la IA para mi archivo?
De la teoría a la práctica

Análisis y extracción de metadatos en contenidos audiovisuales

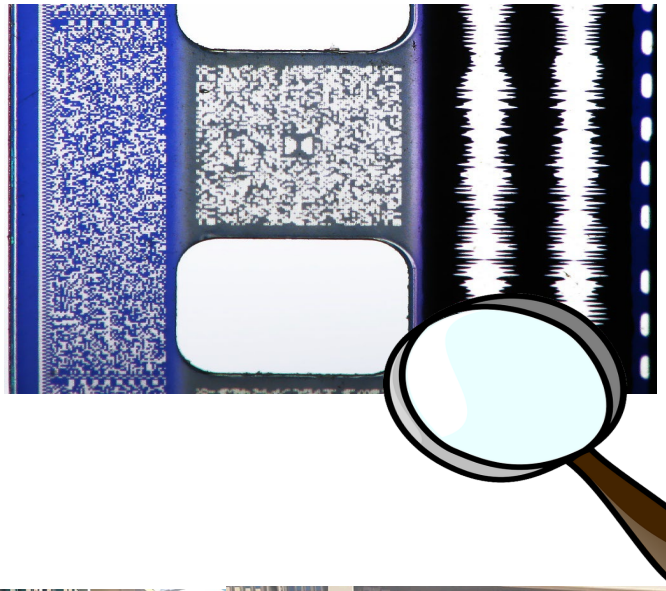
Tecnologías del Habla

Catalogación Asistida



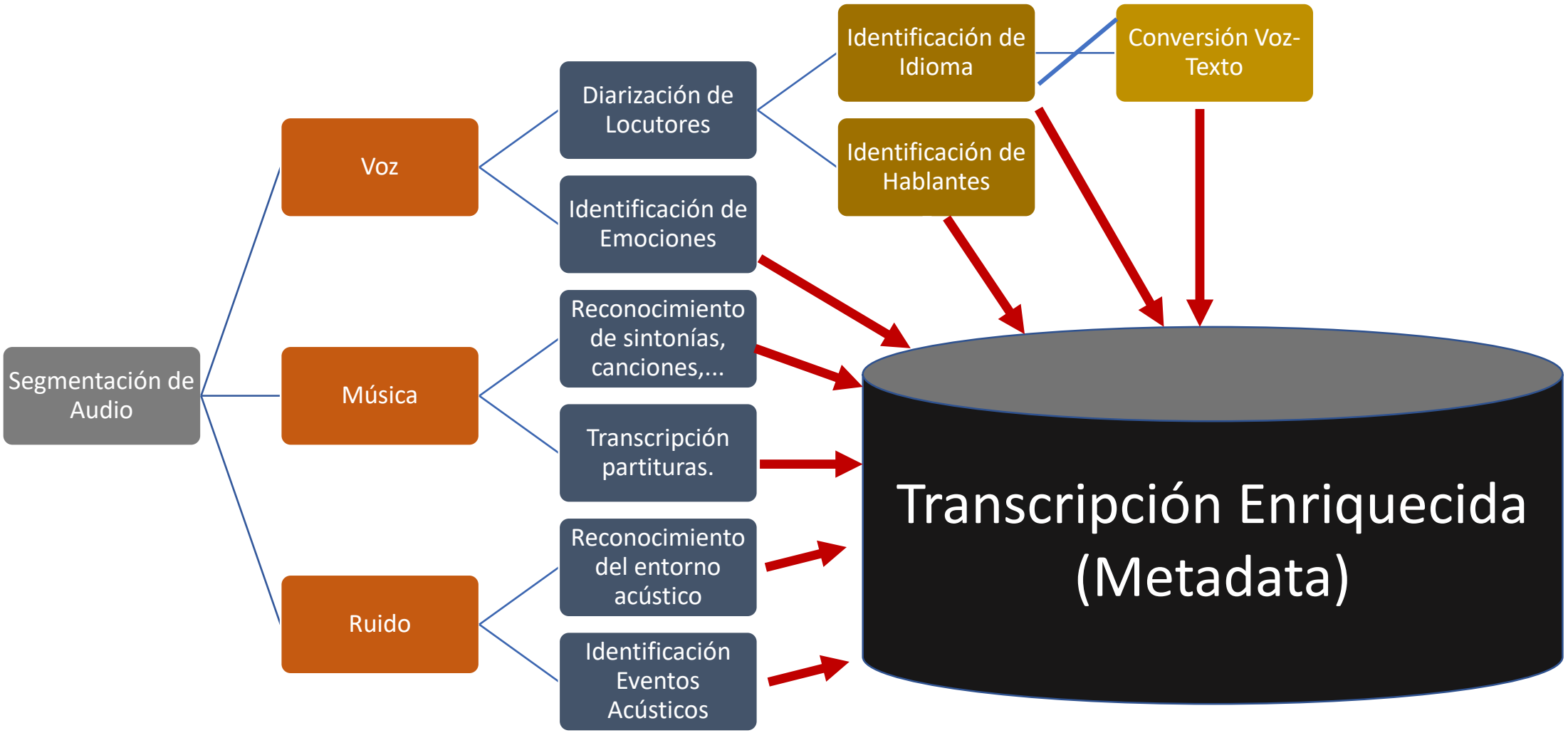


¿Qué información podemos encontrar en un audio?

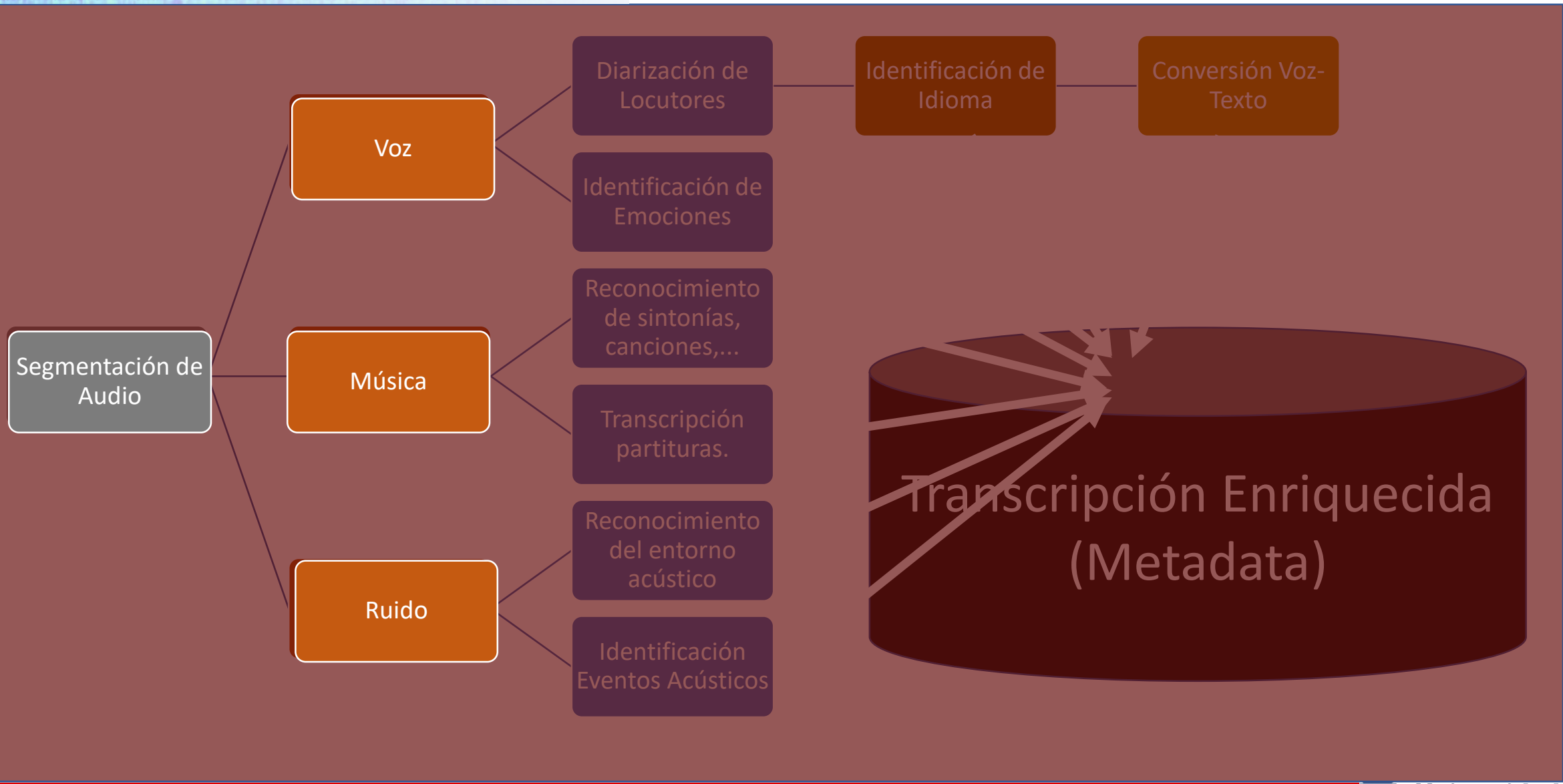


- Hay ruido, música, habla, …
- Cuántas personas hablan y cuándo habla cada una de ellas
- Cuáles son las identidades de las personas que hablan
- En qué idioma están hablando
- Qué dice cada una de ellas
- Cuál es el estado emocional de cada una de ellas

Tecnologías de 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 de Audio

- Herramienta Open-Source
  - inaSpeechSegmenter:

README MIT license

inaSpeechSegmenter

python 3.8 | 3.9 | 3.10 | 3.11 | 3.12 | 3.13 | Python 3.8 to 3.13 unit test passing | pypi package 0.8.0 | docker pulls 537

inaSpeechSegmenter is a CNN-based audio segmentation toolkit suited to the tasks of Voice Activity Detection and Speaker Gender Segmentation.

It splits audio signals into homogeneous zones of speech, music and noise. Speech zones are split into segments tagged using speaker gender (male or female). Male and female classification models are optimized for French language since they were trained using French speakers (acoustic correlates of speaker gender are language dependent). Zones corresponding to speech over music or speech over noise are tagged as speech. Singing voice is tagged as music.

Highlights

- 🏆 ranked #1 against 6 open-source VAD system on a French TV and radio benchmark
- 🏳️‍⚧️ Extended as a non-binary Voice Gender Prediction system for evaluating Transgender voice transition
- 😎 Used since 2020 in the annual French Audivisual Regulation Authority report on women representation in TV and radio!
- 🌟 Used to investigate the relationship between manual and automatic women representation descriptors in French TV and radio
- ♀ ♂ Applied for large-scale gender representation studies in French TV and radio.
- 🥳 Won MIREX 2018 speech detection challenge.

ina Institut national de l'audiovisuel

Command-Line Interface

Binary program ina_speech_segmenter.py may be used to segment multimedia archives encoded in any format supported by ffmpeg. It requires input media and provide 2 segmentation output formats : csv (can be displayed with Sonic Visualiser) and TextGrid (Praat format). Detailed command line options can be obtained using the following command :

```
# get help
$ ina_speech_segmenter.py --help
```

Application Programming Interface

InaSpeechSegmentation API is intended to be very simple to use, and is illustrated by these 2 notebooks :

- Google colab tutorial: use API online
- Jupyter notebook tutorial : to be used offline

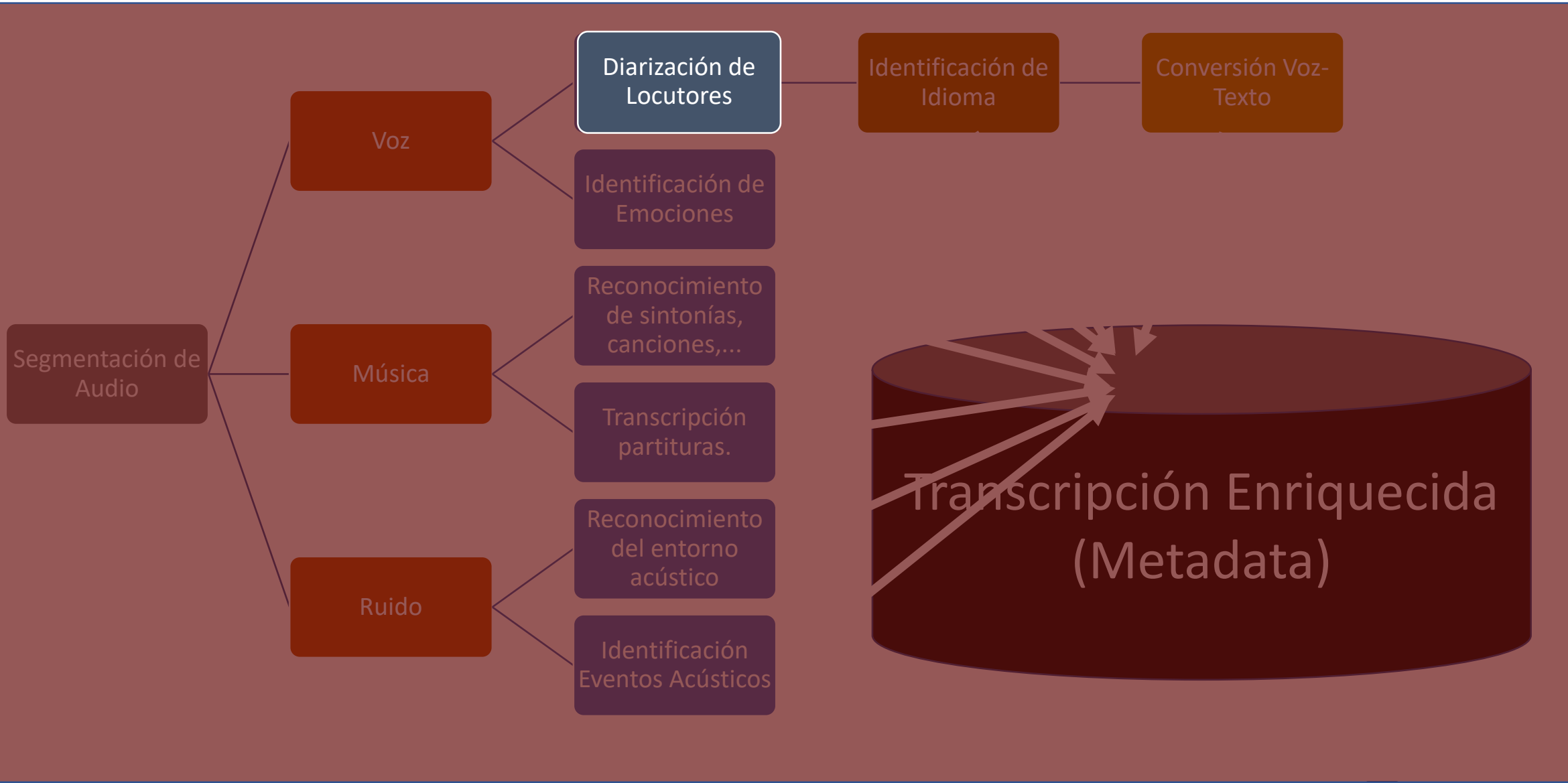
The class allowing to perform segmentations is called Segmenter. It is the only class that you need to import in a program. Class constructor accept 3 optional arguments:

- vad_engine (default: 'smn'). Allows to choose between 2 voice activity detection engines.
  - 'smn' is the more recent engine and splits signal into speech, music and noise segments
  - 'sm' was not trained with noise examples, and split signal into speech and music segments. Noise segments are either considered as speech or music. This engine was used in ICASSP study, and won MIREX 2018 speech detection challenge.
- detect_gender (default: True): if set to True, performs gender segmentation on speech segment and outputs labels 'female' or 'male'. Otherwise, outputs labels 'speech' (faster).
- ffmpeg: allows to provide a specific binary of ffmpeg instead of default system installation

ina

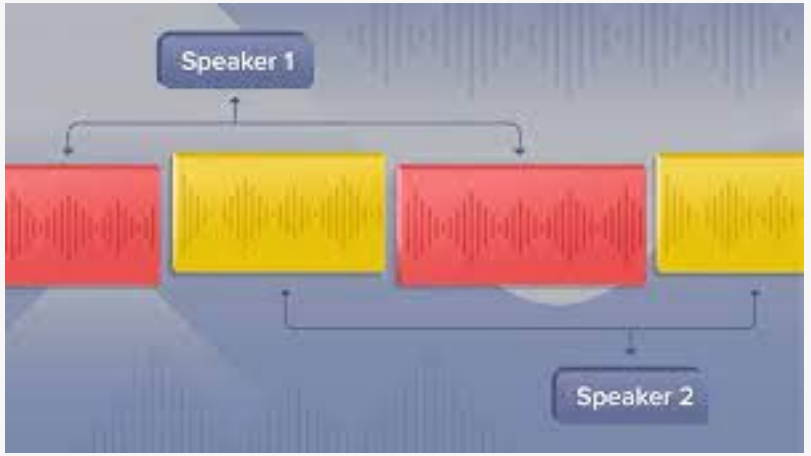
https://github.com/ina-foss/inaSpeechSegmenter

Segmentación y Agrupación de Hablantes



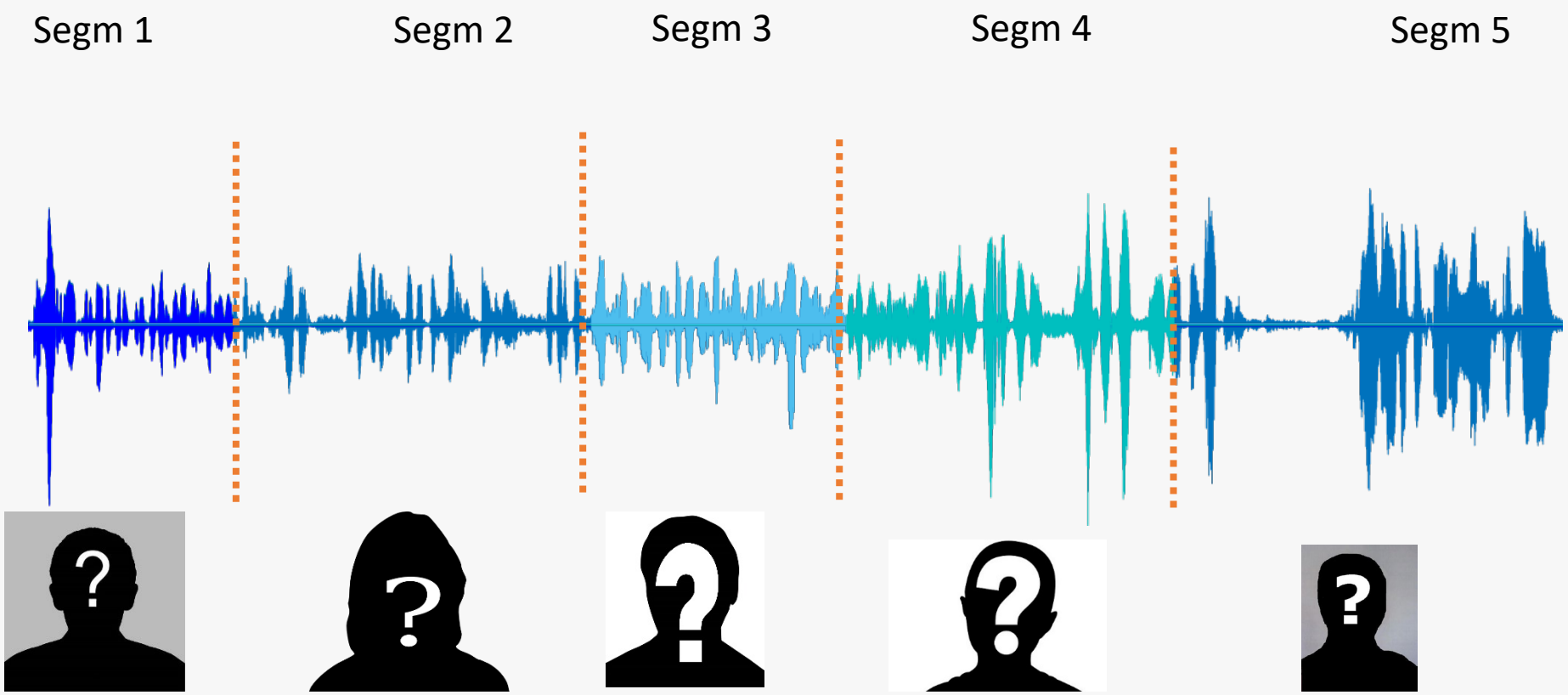
Segmentación y Agrupación de Hablantes

- **¿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**
    - **…**

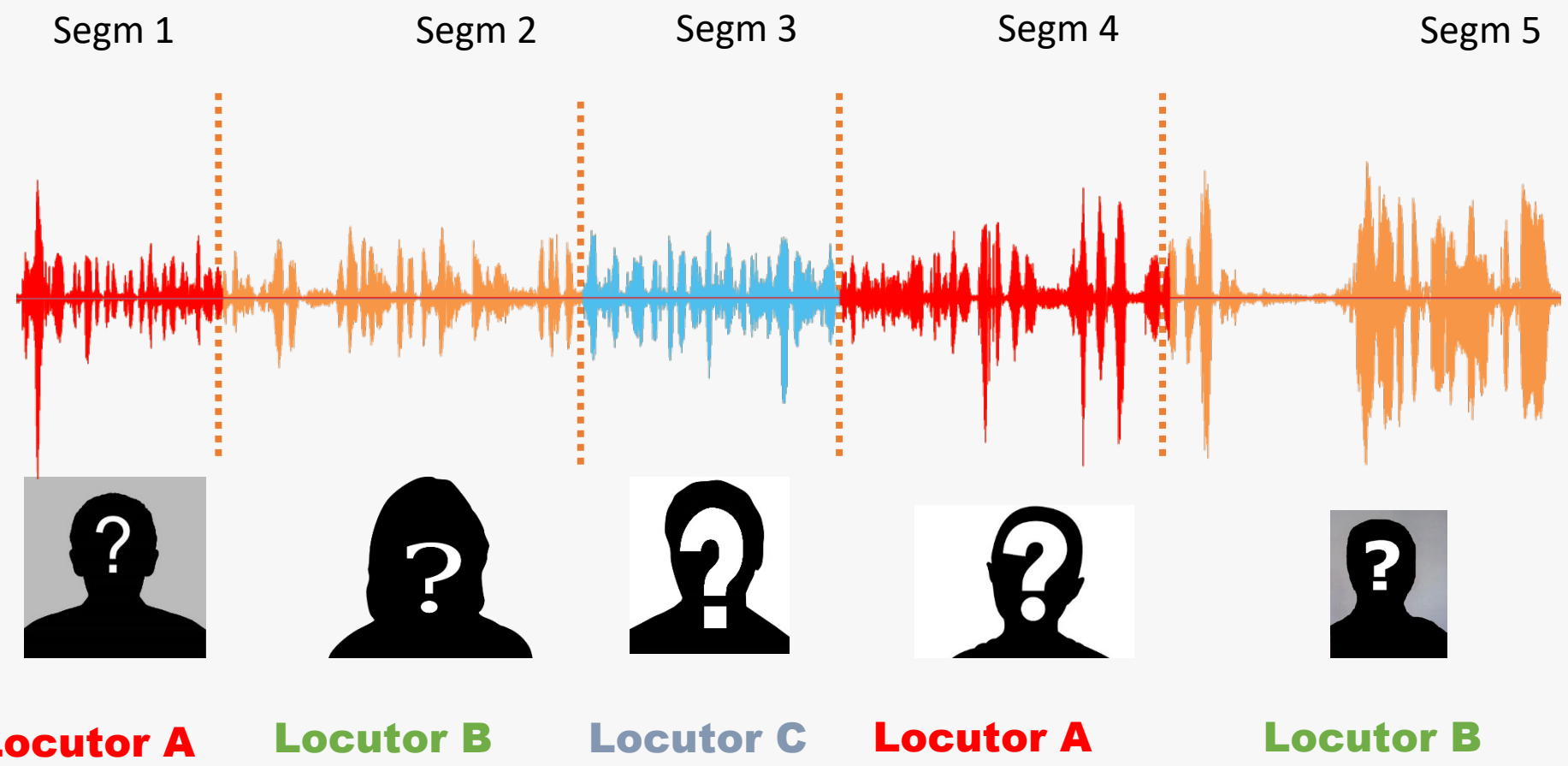


Segmentación y Agrupación de Hablantes

Segmentación y Agrupación de Hablantes

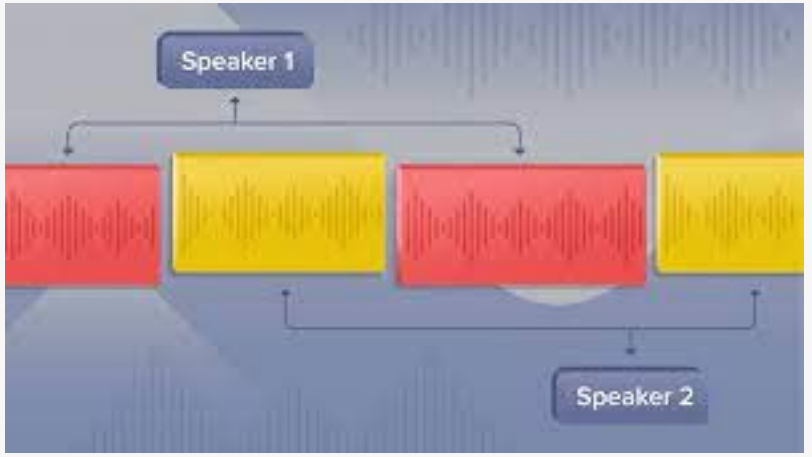


Segmentación y Agrupación de Hablantes



Segmentación y Agrupación de Hablantes

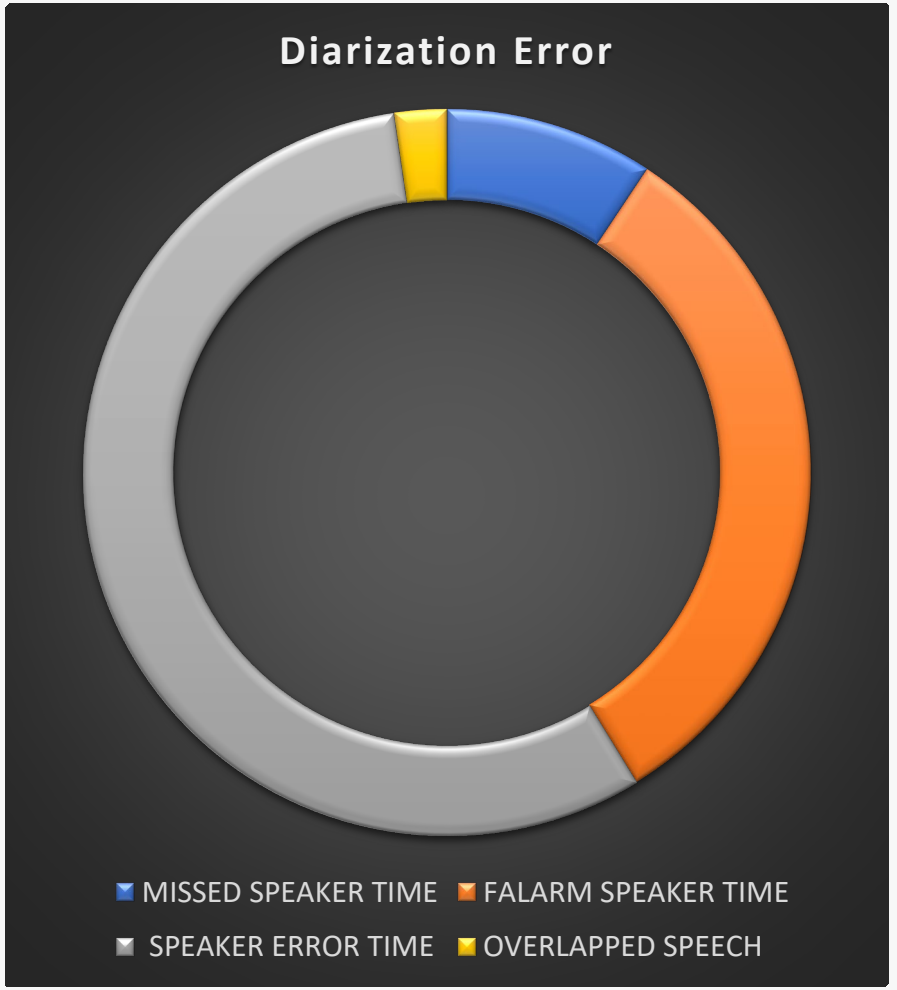
- **¿Para qué sirve?:**
  - **Tecnología soporte para mejorar prestaciones de:**
    - **Reconocimiento automático del habla**
    - **Reconocimiento del hablante**
    - **…**



Medida del Error

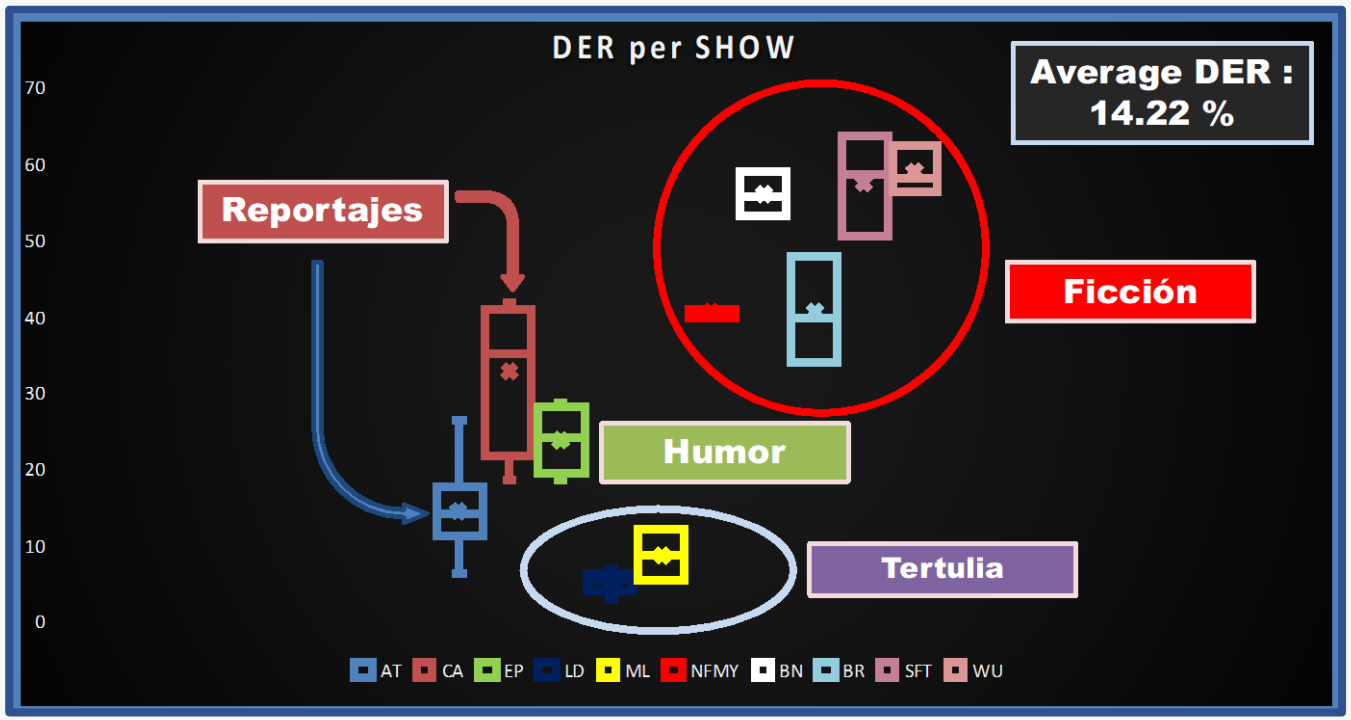
- **Diarization Error Rate:**

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



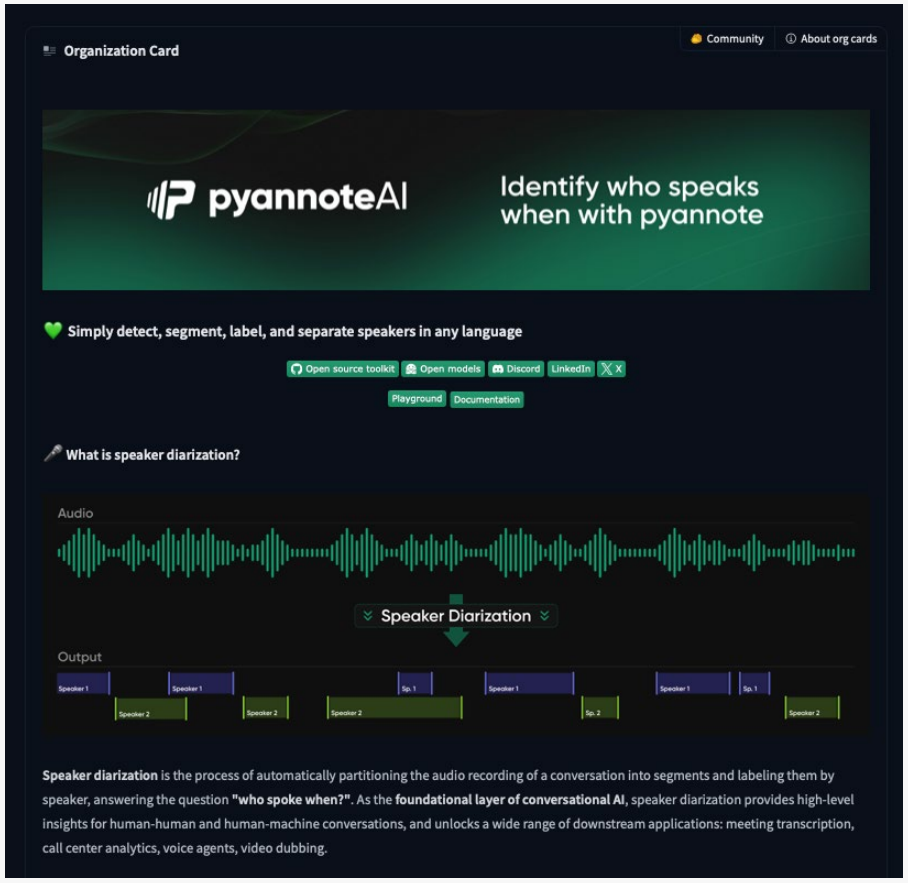
- Componentes del error:
  - Pérdida:
    - Hay voz pero se ha confundido con silencio.
  - Falsa Alarma:
    - No hay voz pero se ha detectado erróneamente.
  - Error de Locutor:
    - Se ha confundido un locutor con otro.
  - Error por Solape:
    - Dos locutores hablan a la vez, pero solo se ha identificado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

Prestaciones: Albayzín Retos - RTVE

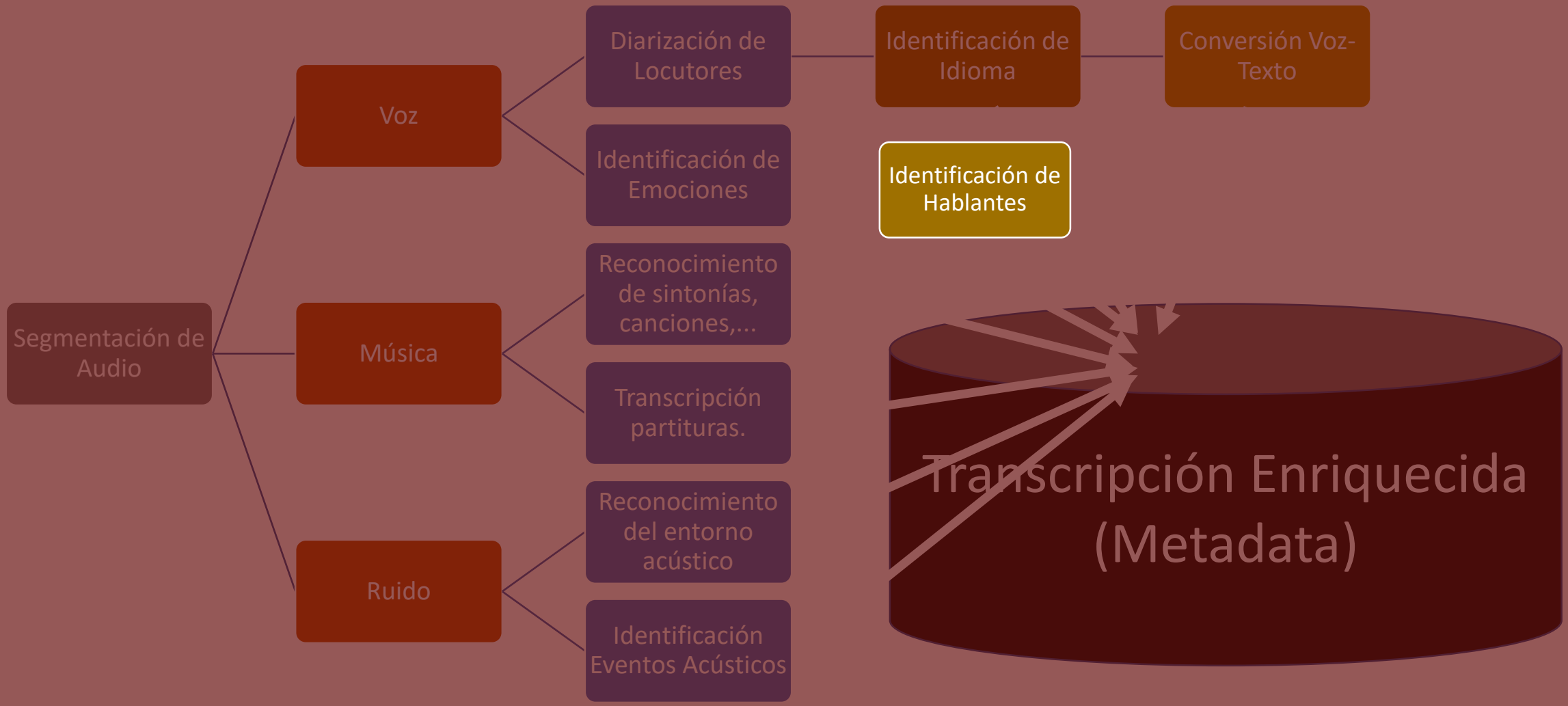


Segmentación y Agrupación de Hablantes

- <u>Herramienta Open-Source</u>
  - Pyannote:



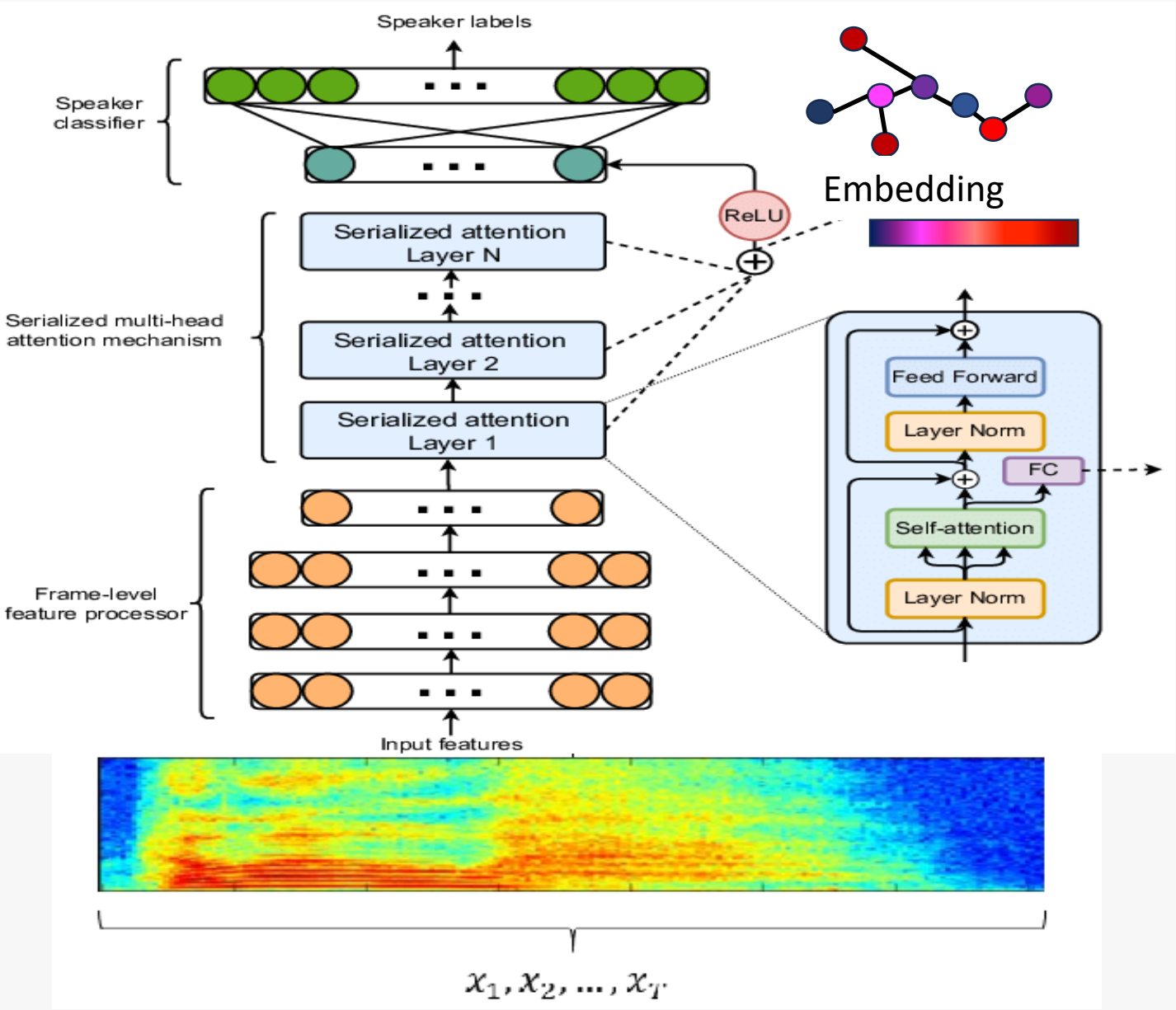
Identificación de Hablantes



Identificación de Hablantes

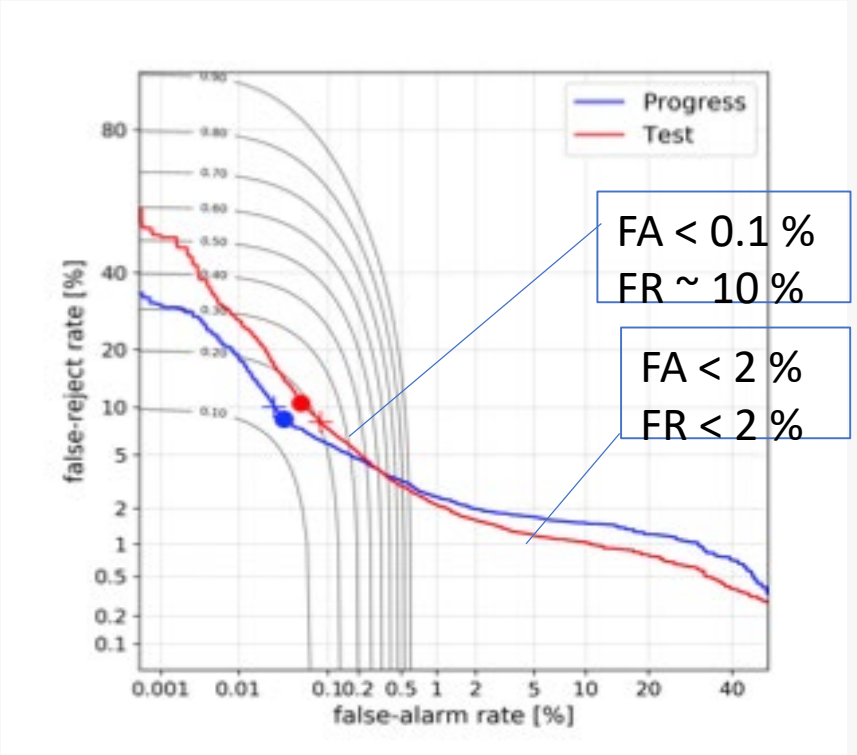
- **¿Para qué sirve?:**
  - **Permite asignar identidades concretas a fragmentos de audio de un contenido analizado**

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

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, …**

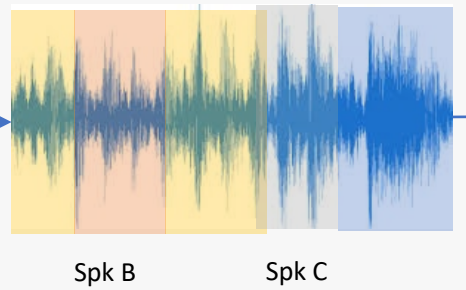
- **Reto para el Archivo: Variabilidad Intra-Locutor**
  - El mismo hablante suena distinto según el contexto.
    - **Envejecimiento**:
      La voz de un político en 1980 y su voz en 2010 no son iguales.
    - **Entorno**:
      El mismo periodista hablando en un estudio insonorizado (voz limpia, grave) o informando desde una zona en guerra (gritando, micrófono de mano, ruido).

Identificación de Hablantes

- Reconocimiento de personas en programas de TV:

Extraer Audio

Diarizar Audio



Asignar Identidad Hablante

Motor Reconocimiento Biométrico

Registrar Locutores

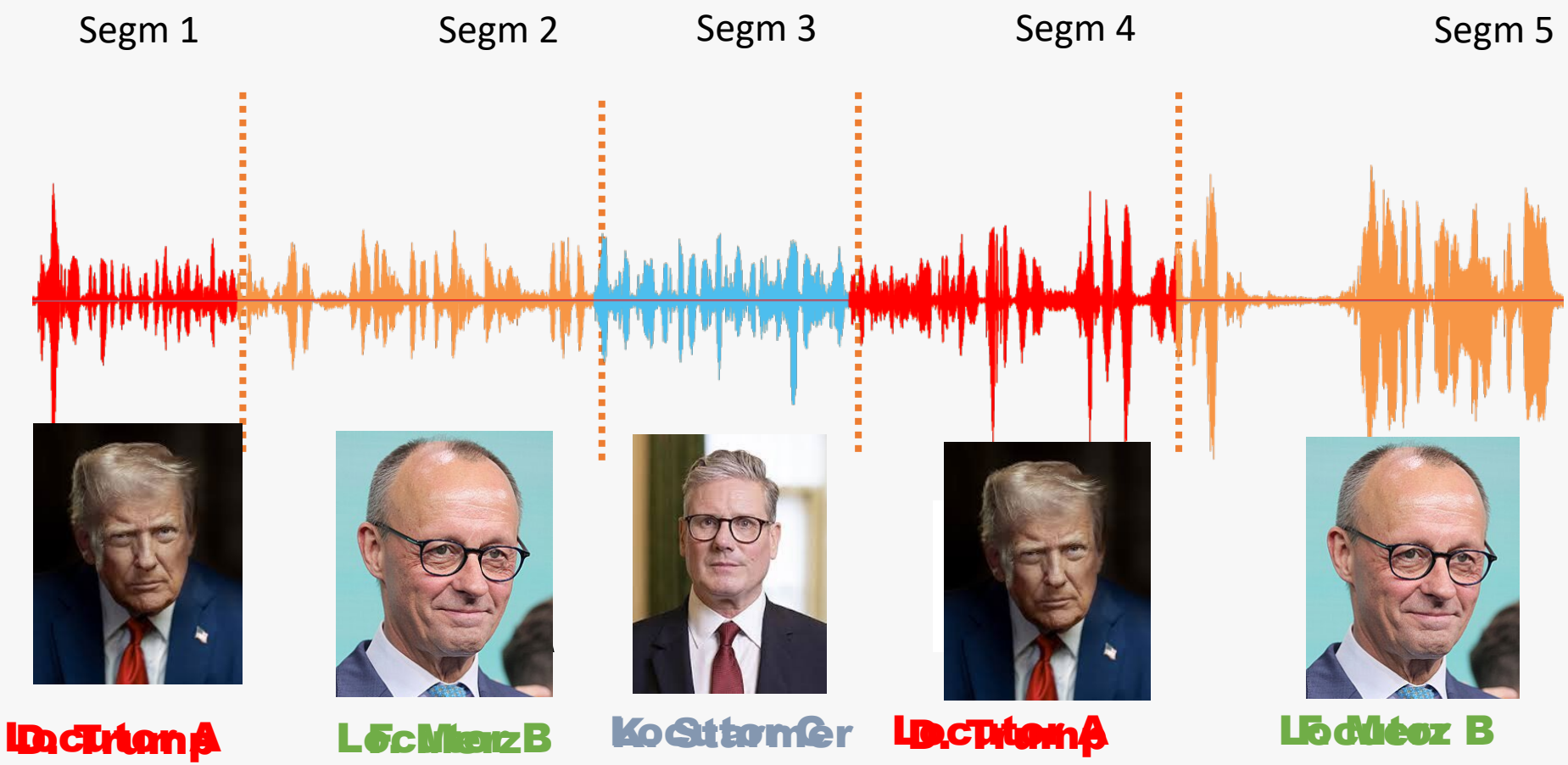
Base de Datos Hablantes

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

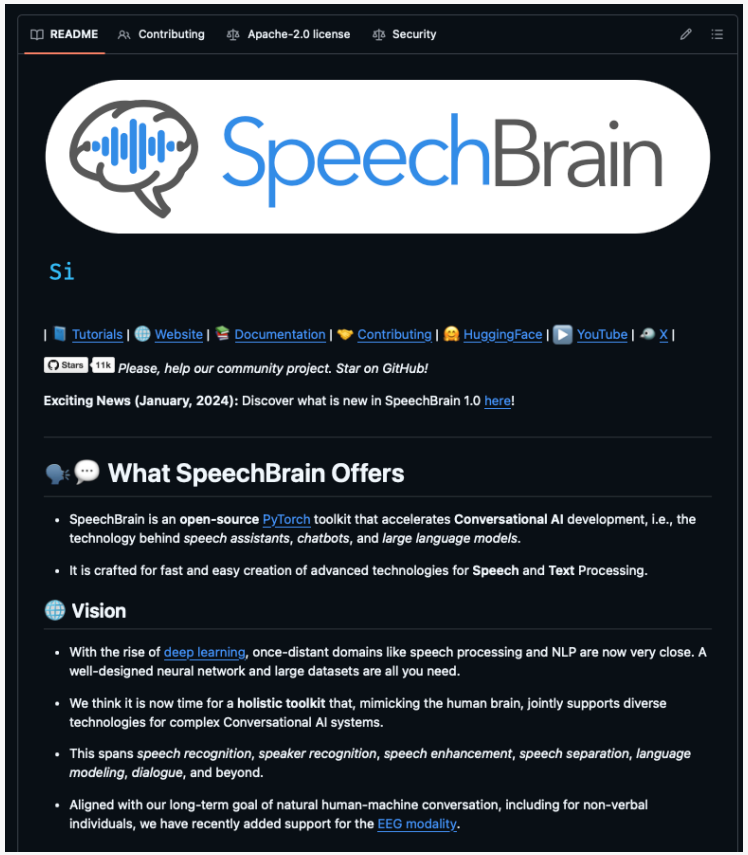
	AER
BIOMETRIC VOX	65.09 %
VIVOLAB	72.63 %

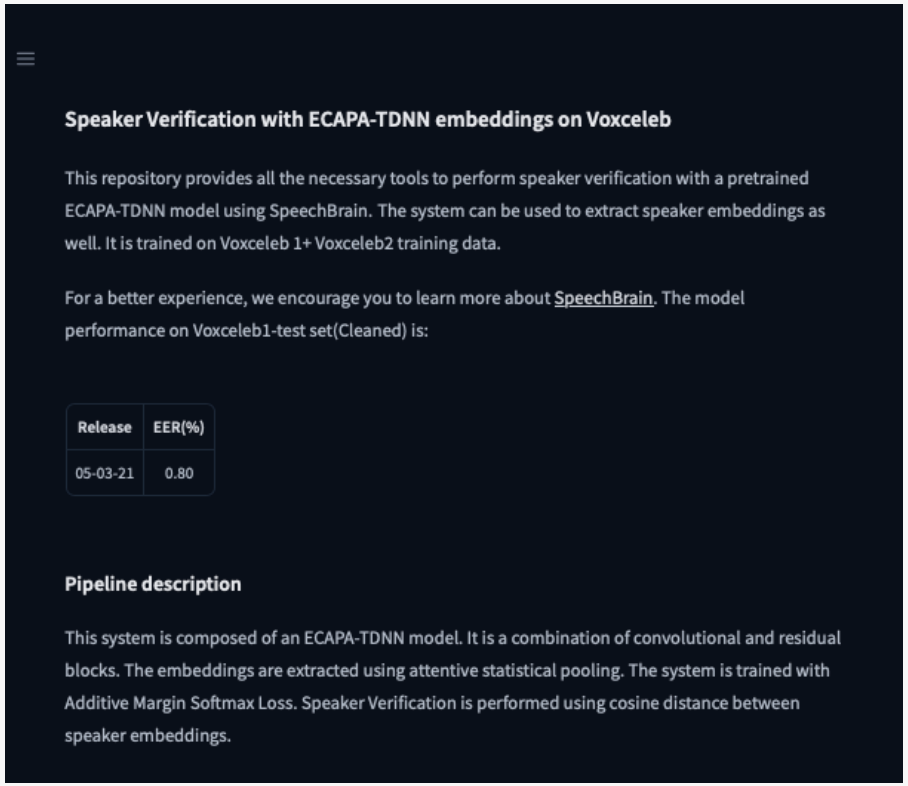
	MISS	FA	SPKERR	AER
TEAMIV_p	1,3	229,7	9,5	240,55
TEAMIV_l3	3,7	160,8	20,9	185,42
TEAMIV_l6	8,3	76,5	5,9	90,62
TEAMIV_l9	**12,4**	**15,3**	**1,2**	**28,88**

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 Hablantes

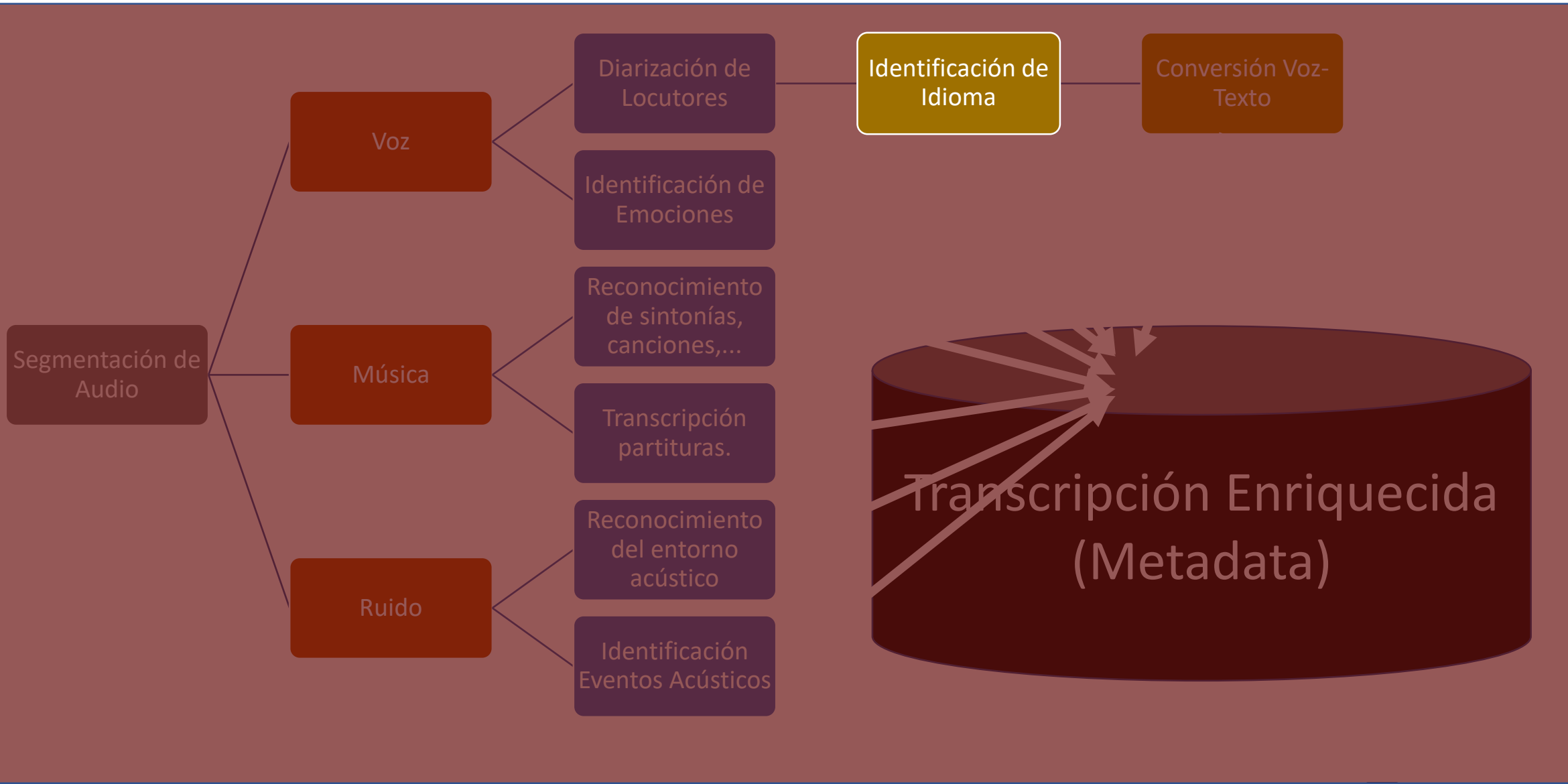
- **Herramienta Open-Source**
  - **SpeechBrain:**





https://github.com/speechbrain/speechbrain

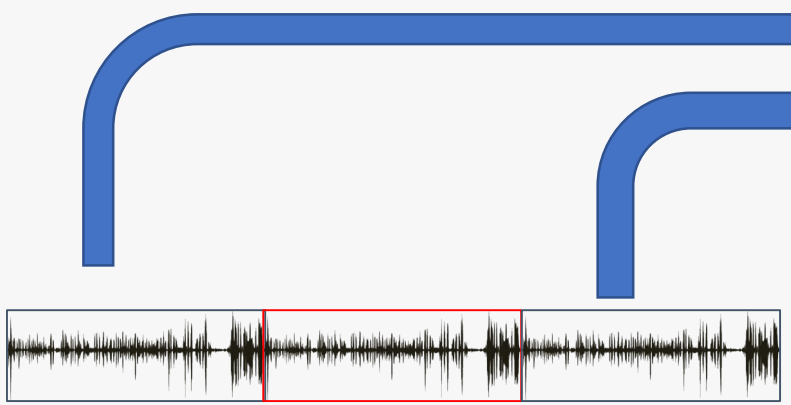
Tecnologías



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)

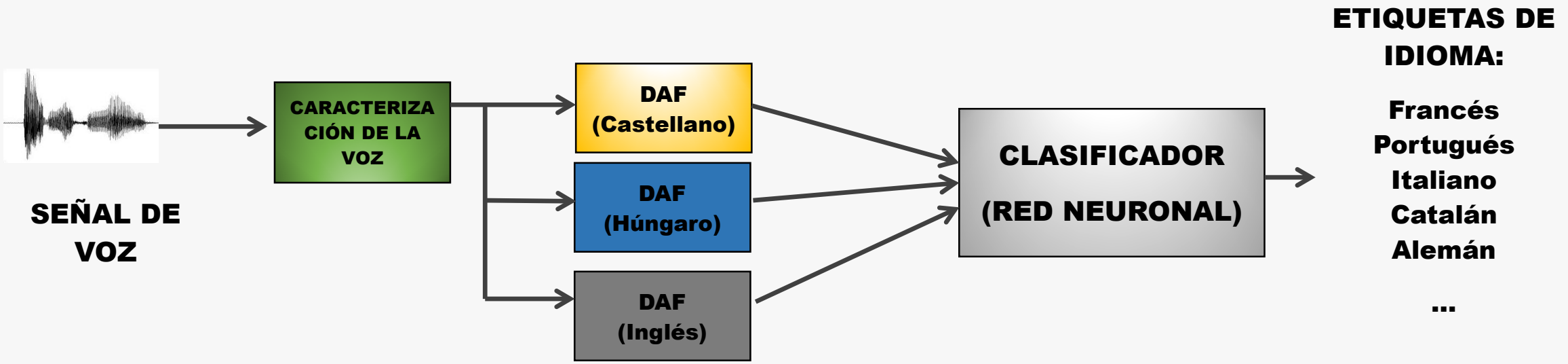
La Fundación ”la Caixa” ha hecho balance de la asistencia a sus centros y exposiciones culturales y científicas durante 2021, año todavía marcado por la pandemia. La entidad ha hecho un gran esfuerzo para recuperar el ritmo de visitantes en los ocho centros CaixaForum y en CosmoCaixa, y en las exposiciones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

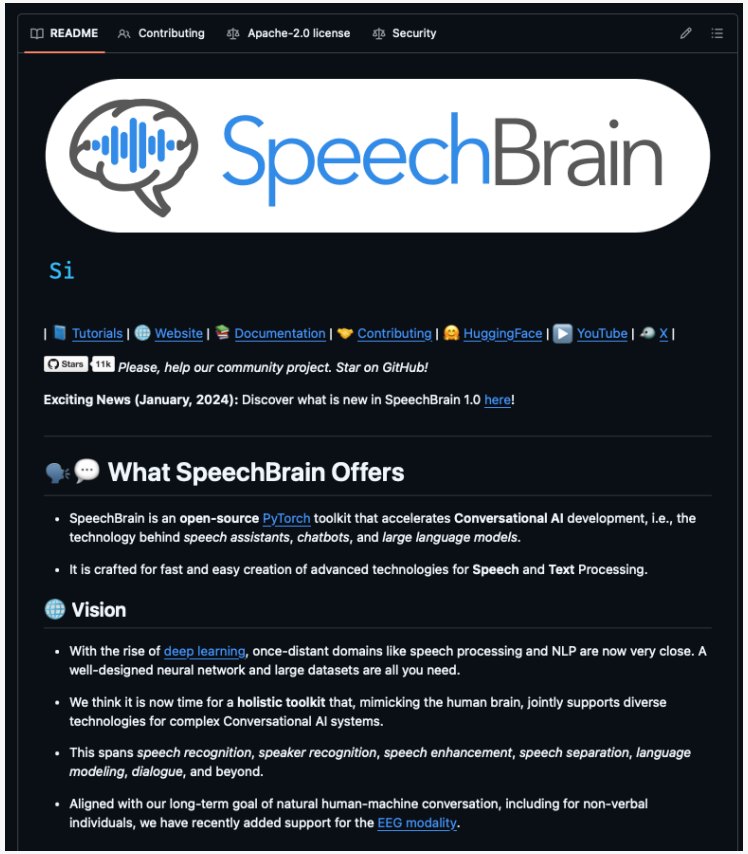
- **Acústicos**
  - Tratan de buscar patrones discriminativos directamente sobre la señal de voz
- **Fonotácticos (Lingüísticos)**
  - Primero procesan la señal de entrada con un (o varios) reconocedor fonético (en varios idiomas) después buscan patrones discriminativos en las secuencias de fonemas de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

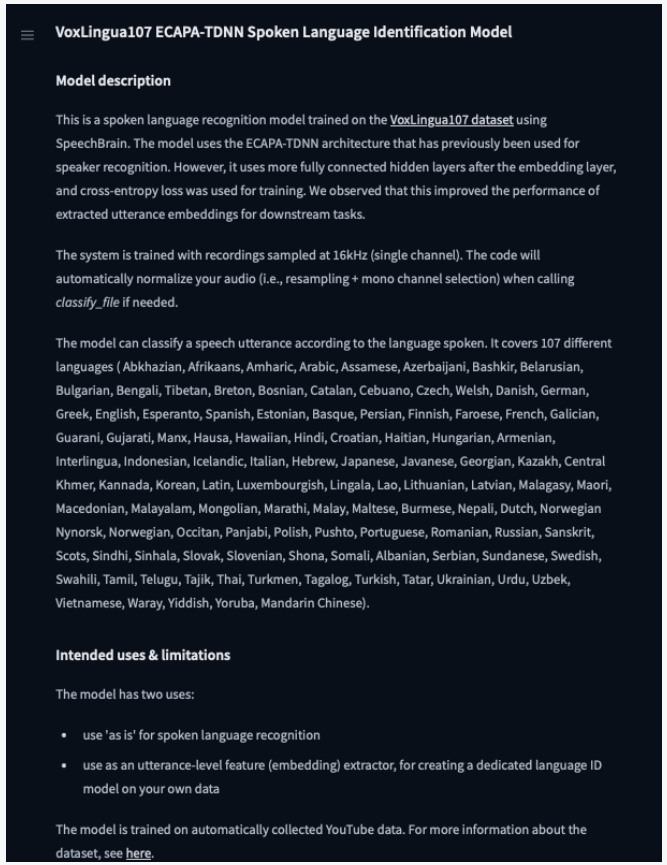


Identificación de Idioma

- **Herramienta Open-Source**

  - **SpeechBrain:**

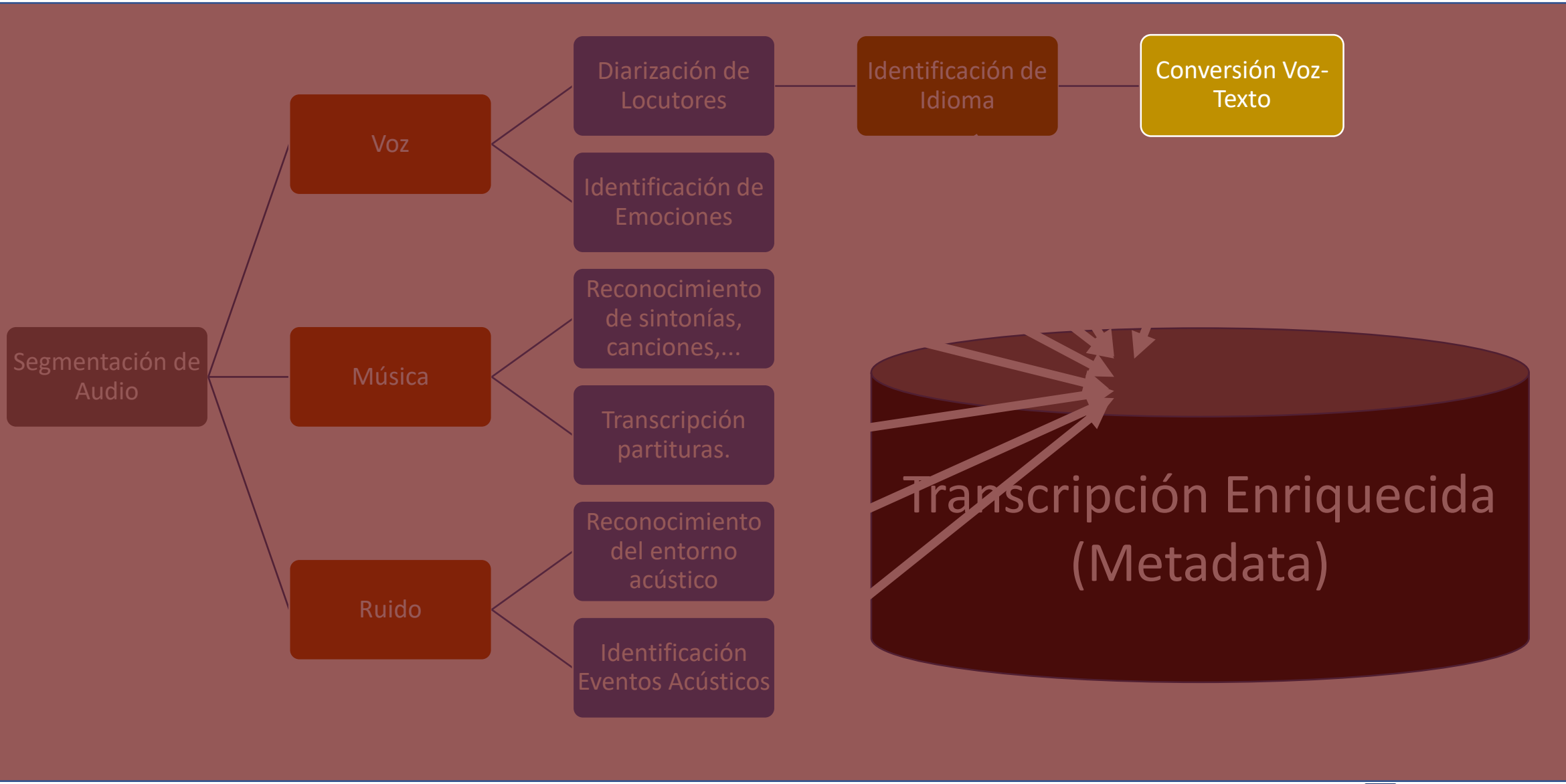




https://github.com/speechbrain/speechbrain

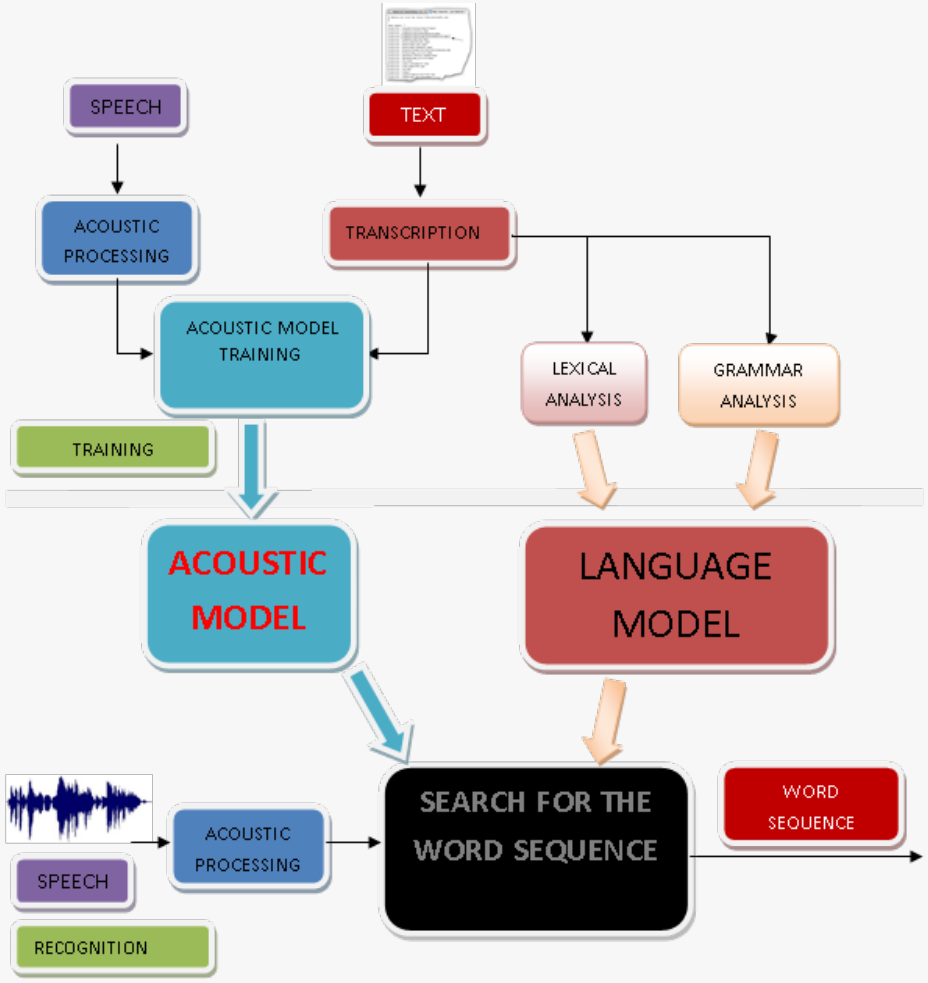
https://huggingface.co/speechbrain/lang-id-voxlingua107-ecapa

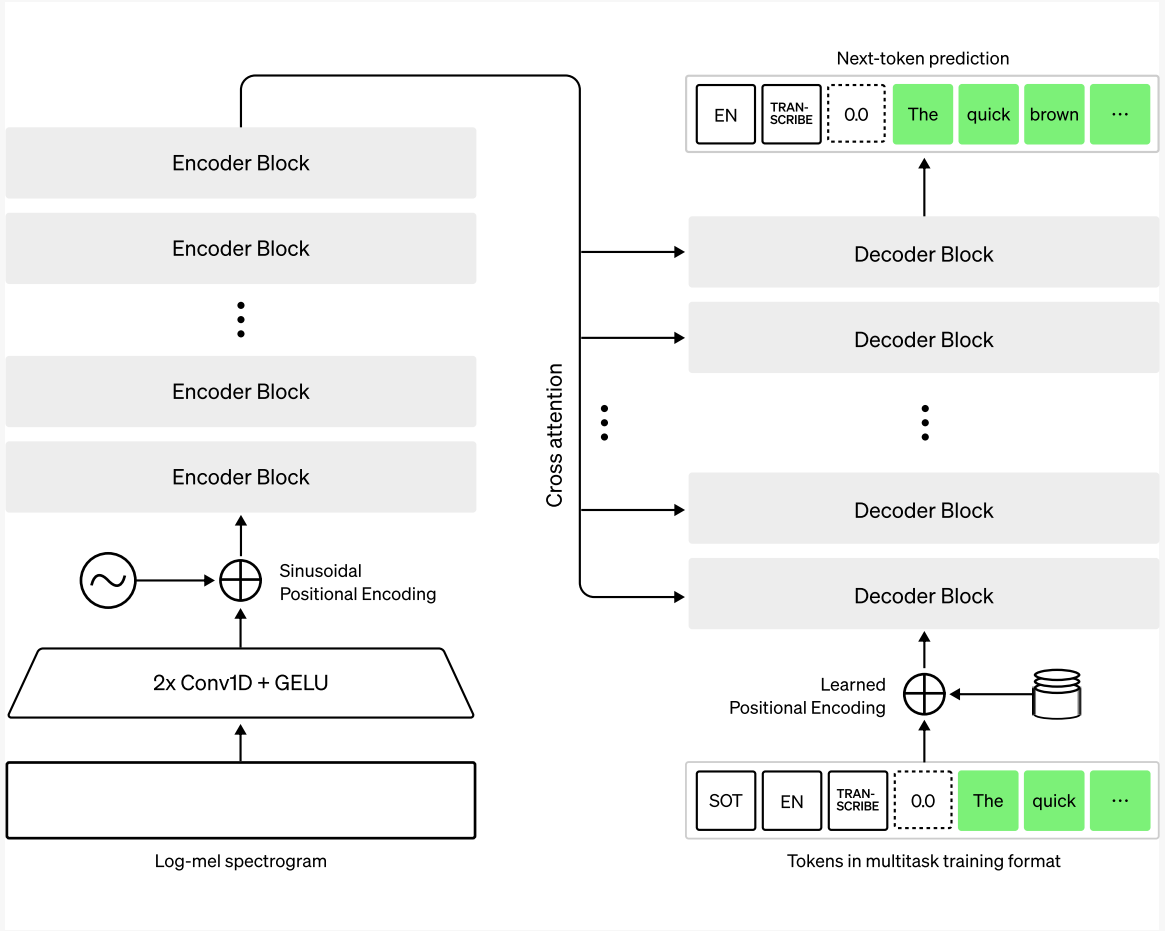
Tecnologías



Reconocimiento Automático del Habla (ASR):

- Transforma un documento audiovisual en algo **CONSULTABLE**





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>

Reconocimiento Automático del Habla (ASR):

- **Herramienta Open-Source**

  - **Whisper:**

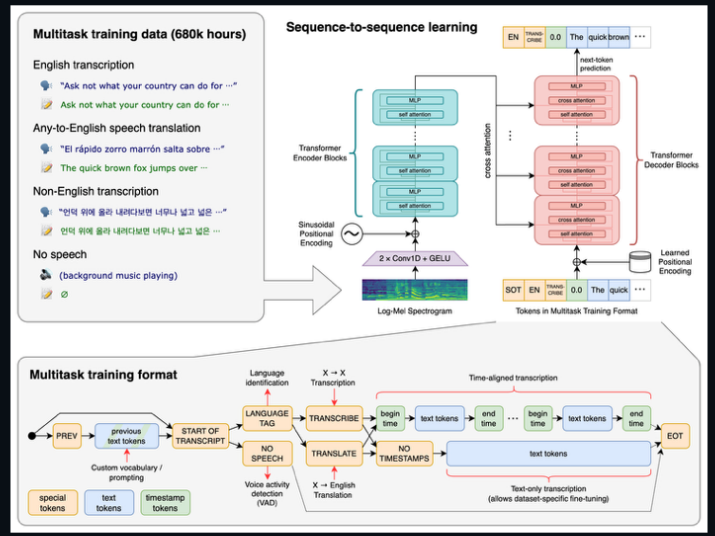
README MIT license

Whisper

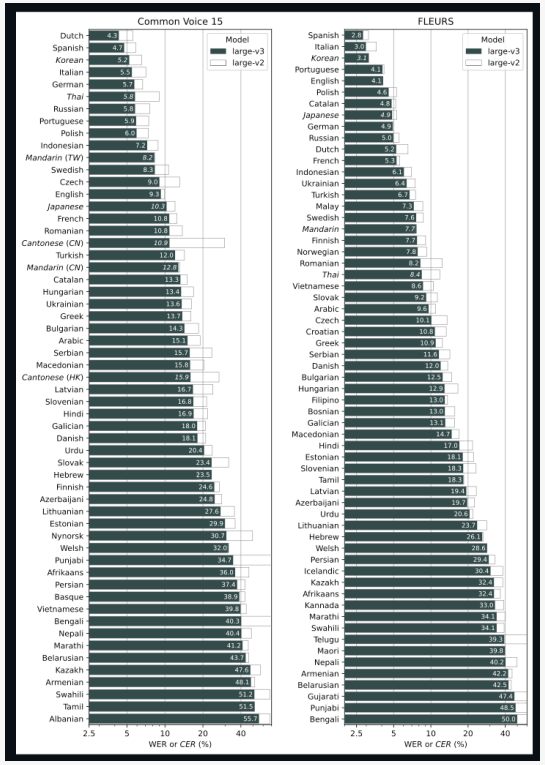
[Blog] [Paper] [Model card] [Colab example]

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification.

Approach



A Transformer sequence-to-sequence model is trained on various speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. These tasks are jointly represented as a sequence of tokens to be predicted by the decoder, allowing a single model to replace many stages of a traditional speech-processing pipeline. The multitask training format uses a set of special tokens that serve as task specifiers or classification targets.



Whisper

Whisper is a state-of-the-art model for automatic speech recognition (ASR) and speech translation, proposed in the paper Robust Speech Recognition via Large-Scale Weak Supervision by Alec Radford et al. from OpenAI. Trained on >5M hours of labeled data, Whisper demonstrates a strong ability to generalise to many datasets and domains in a zero-shot setting.

Whisper large-v3 has the same architecture as the previous large and large-v2 models, except for the following minor differences:

1. The spectrogram input uses 128 Mel frequency bins instead of 80
2. A new language token for Cantonese

The Whisper large-v3 model was trained on 1 million hours of weakly labeled audio and 4 million hours of pseudo-labeled audio collected using Whisper large-v2 . The model was trained for 2.0 epochs over this mixture dataset.

The large-v3 model shows improved performance over a wide variety of languages, showing 10% to 20% reduction of errors compared to Whisper large-v2 . For more details on the different checkpoints available, refer to the section Model details.

Disclaimer: Content for this model card has partly been written by the 🤗 Hugging Face team, and partly copied and pasted from the original model card.

Usage

https://github.com/openai/whisper
https://huggingface.co/openai/whisper-large-v3

Reconocimiento Automático del Habla (ASR):

- **Herramienta Open-Source**
  - **WhisperX:**
    - Alineamiento Forzado (Forced Alignment)
      - Toma un texto ya transcrito y fuerza a coincidir con su audio, sonido a sonido
      - Precisión de palabra:
        - Da el timestamp exacto de cada palabra.
      - Diarización integrada:
        - WhisperX integra pyannote.audio.
        - Proporciona:
          - Texto + Tiempo Exacto + Quién habla.

VGG
UNIVERSITY OF OXFORD

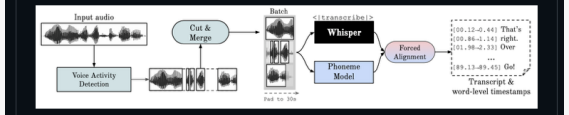
README BSD-2-Clause license

WhisperX

Recall.ai - Meeting Transcription API

If you're looking for a transcription API for meetings, consider checking out Recall.ai's Meeting Transcription API, an API that works with Zoom, Google Meet, Microsoft Teams, and more. Recall.ai diarizes by pulling the speaker data and separate audio streams from the meeting platforms, which means 100% accurate speaker diarization with actual speaker names.

Stars 20k issues 249 open license BSD-2-Clause Arxiv 2303.00747 Tweet



This repository provides fast automatic speech recognition (70x realtime with large-v2) with word-level timestamps and speaker diarization.

- ⚡ Batched inference for 70x realtime transcription using whisper large-v2
- 🪶 faster-whisper backend, requires <8GB gpu memory for large-v2 with beam_size=5
- 🎯 Accurate word-level timestamps using wav2vec2 alignment
- 👯 Multispeaker ASR using speaker diarization from pyannote-audio (speaker ID labels)
- 🗣️ VAD preprocessing, reduces hallucination & batching with no WER degradation

Whisper is an ASR model developed by OpenAI, trained on a large dataset of diverse audio. Whilst it does produces highly accurate transcriptions, the corresponding timestamps are at the utterance-level, not per word, and can be inaccurate by several seconds. OpenAI's whisper does not natively support batching.

Phoneme-Based ASR A suite of models finetuned to recognise the smallest unit of speech distinguishing one word from another, e.g. the element p in "tap". A popular example model is wav2vec2.0.

Forced Alignment refers to the process by which orthographic transcriptions are aligned to audio recordings to automatically generate phone level segmentation.

Voice Activity Detection (VAD) is the detection of the presence or absence of human speech.

Speaker Diarization is the process of partitioning an audio stream containing human speech into homogeneous segments according to the identity of each speaker.

Usage 💬 (command line)

English

Run whisper on example segment (using default params, whisper small) add `--highlight_words True` to visualise word timings in the .srt file.

```
whisperx path/to/audio.wav
```

Result using *WhisperX* with forced alignment to wav2vec2.0 large:

sample01.mp4

Compare this to original whisper out the box, where many transcriptions are out of sync:

sample_whisper_og.mov

Bella, Gi…, love.

For increased timestamp accuracy, at the cost of higher gpu mem, use bigger models (bigger alignment model not found to be that helpful, see paper) e.g.

```
whisperx path/to/audio.wav --model large-v2 --align_model WAV2VEC2_ASR_LARGE_LV60K_960H --b
```

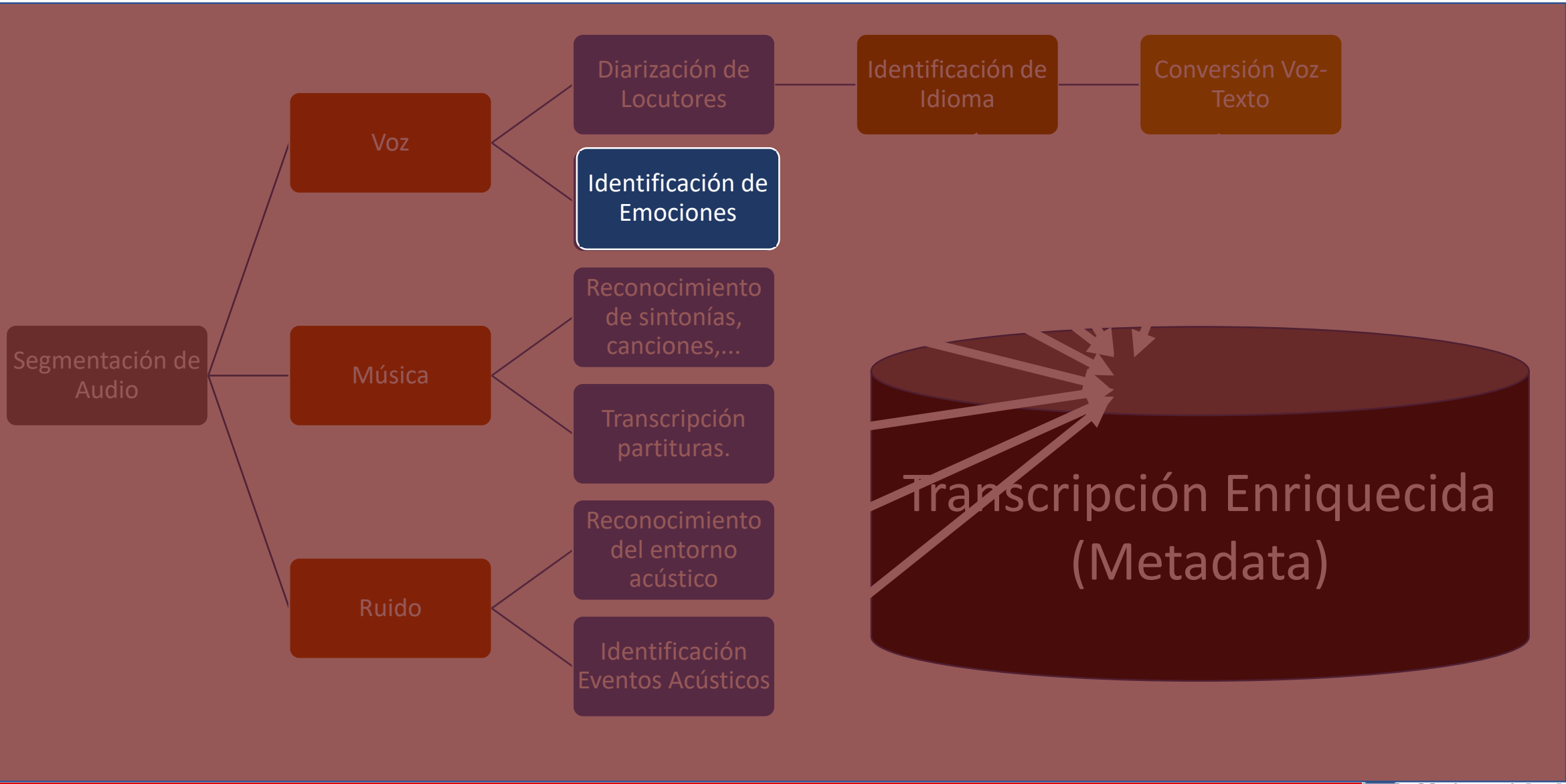
To label the transcript with speaker ID's (set number of speakers if known e.g. `--min_speakers 2 --max_speakers 2`):

```
whisperx path/to/audio.wav --model large-v2 --diarize --highlight_words True
```

VGG
UNIVERSITY OF OXFORD

https://github.com/m-bain/whisperX

Tecnologías

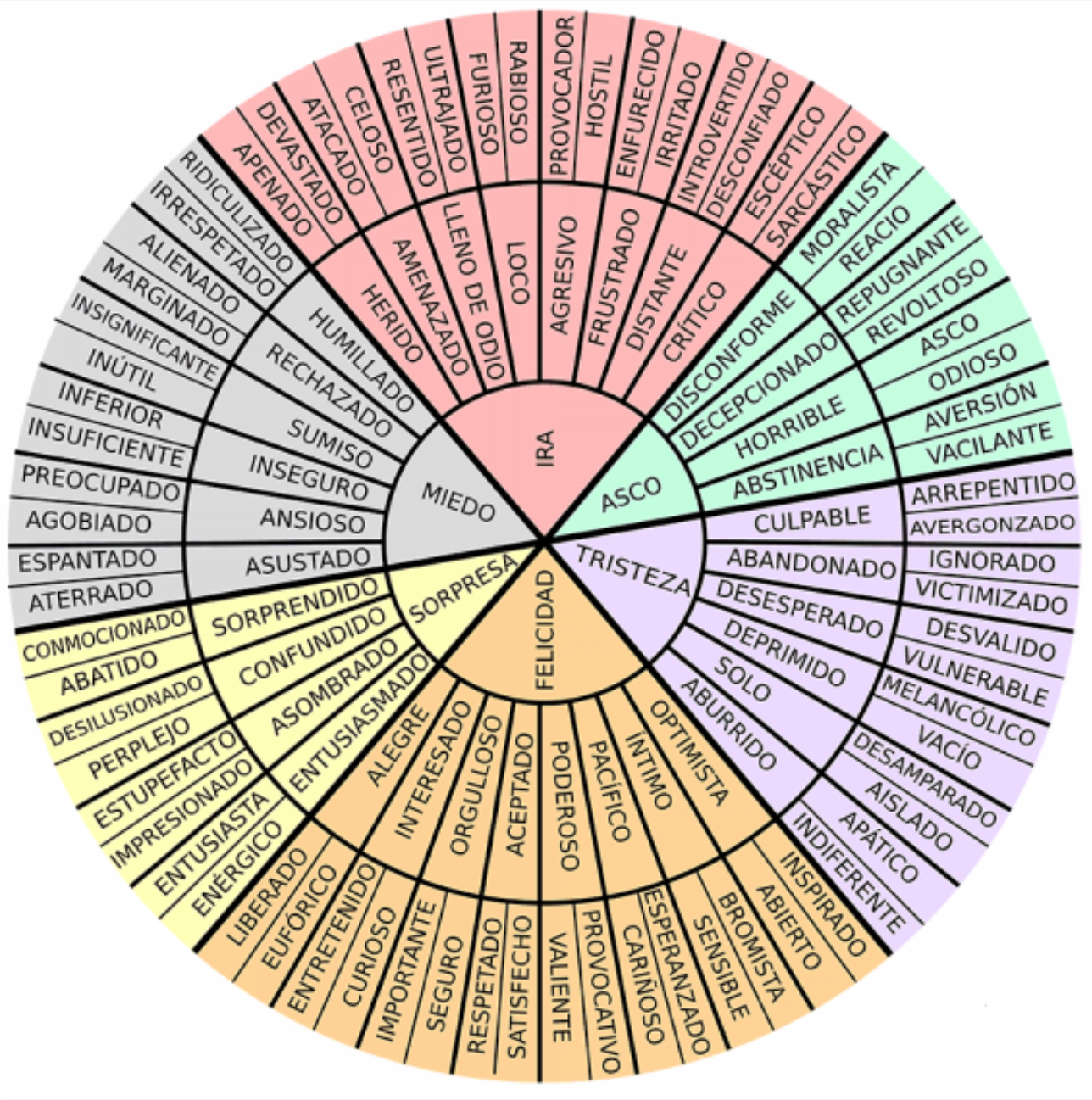


Identificación de Emociones

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

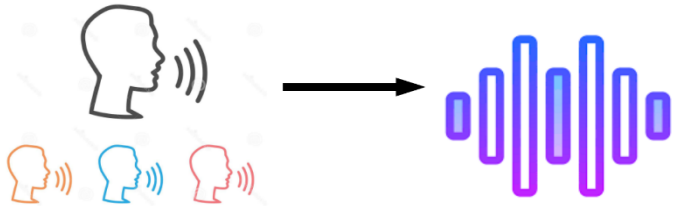
El problema de los datos

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

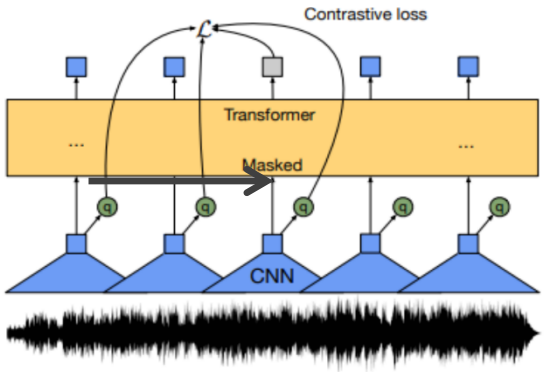


Identificación de Emociones

SEÑAL DE VOZ



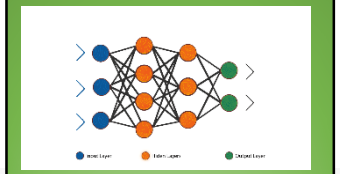
Feature Extraction Using Wav2Vec2



Classify Layer

Feature Vectors

CLASIFICACIÓN



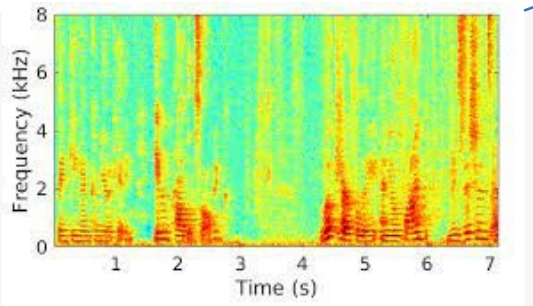
Emotion Recognition

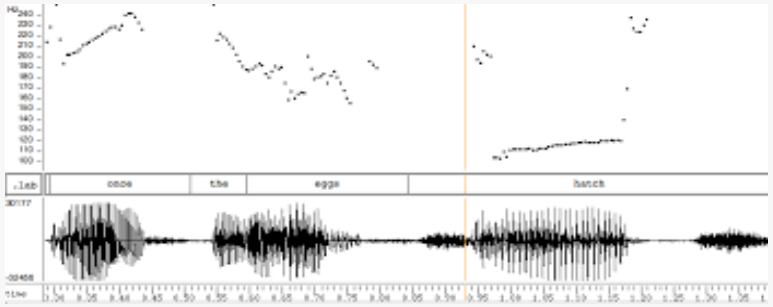
- enfado
- alegría
- tristeza
- neutro
- aburrimiento
- ansiedad

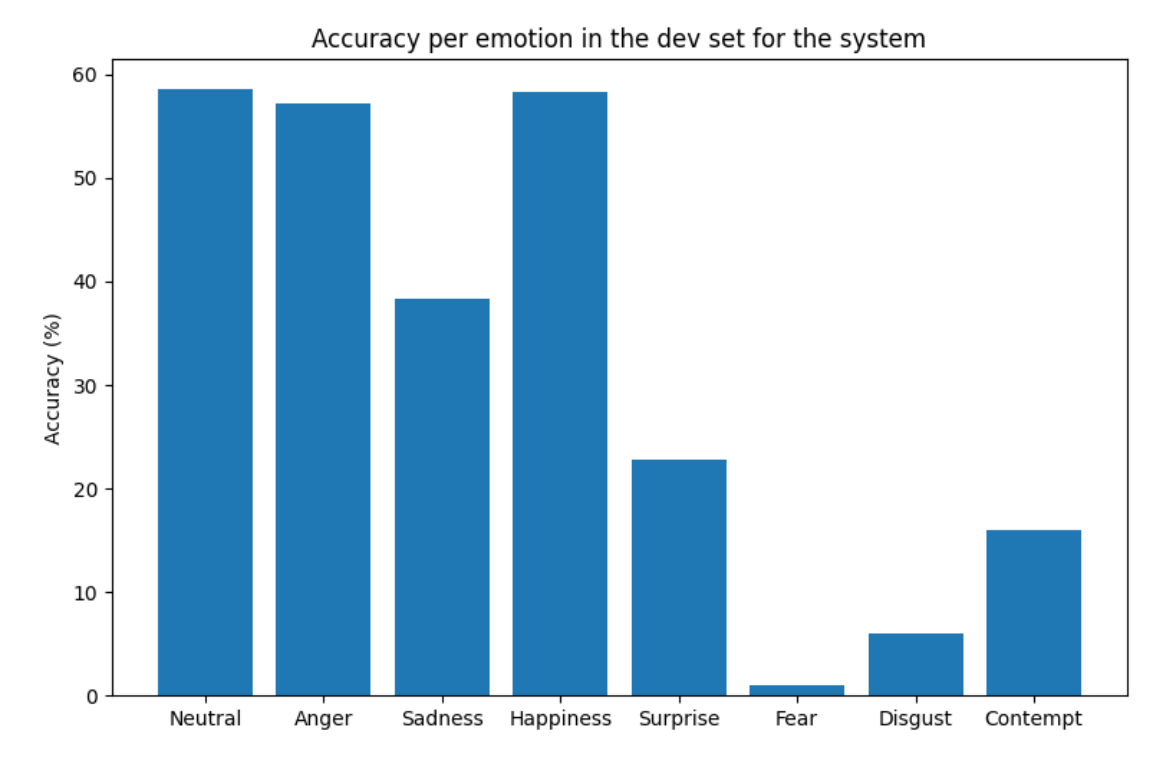
Características:

Espectrales
Prosódicas
Paralingüísticas

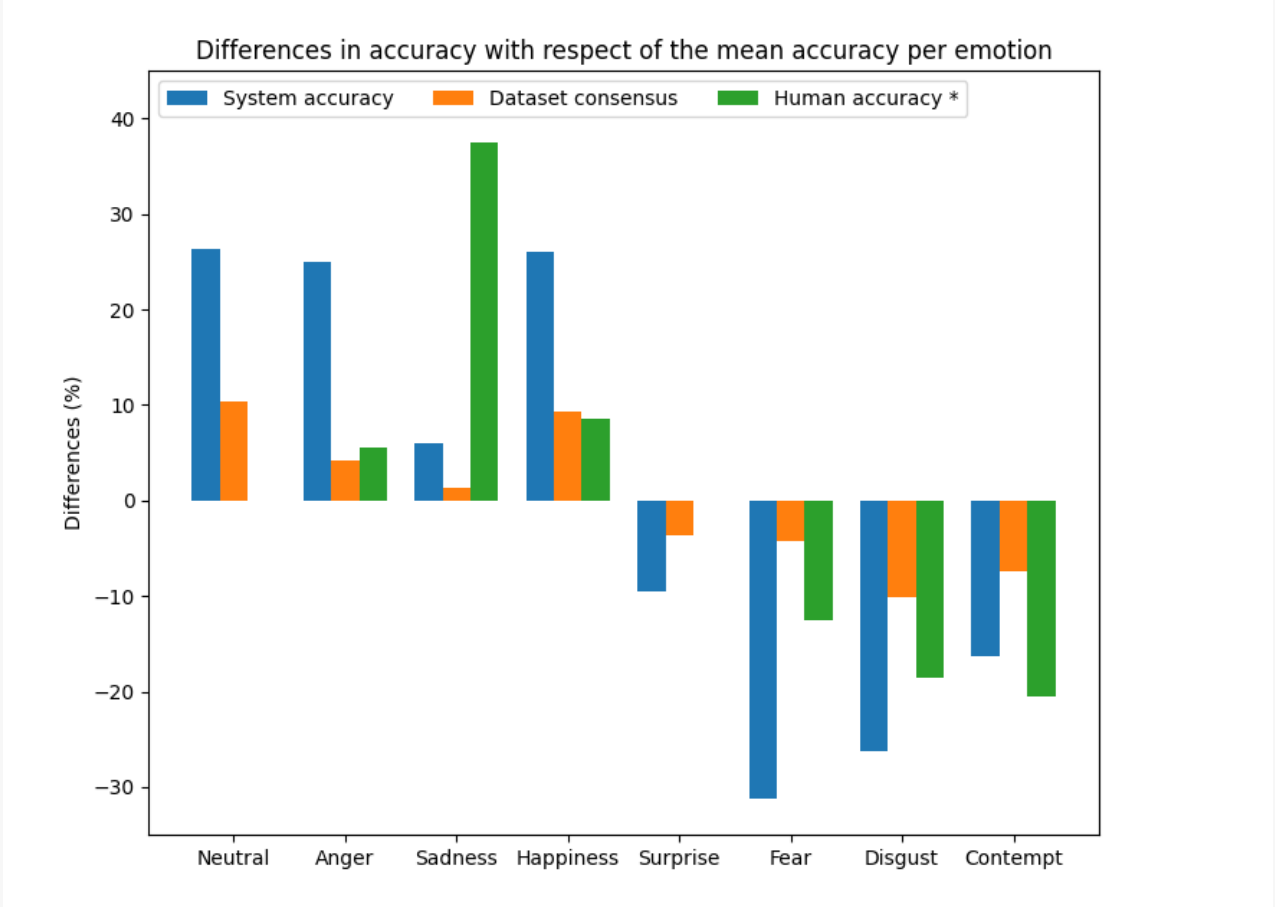
...







- 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.



Restauración de Audios

- Cambio de Paradigma: Sustractivo vs. Generativo
  - **Métodos Tradicionales**:
    - Eliminar las componentes del ruido manteniendo todo lo posible las de la voz
  - **Métodos Actuales** (Generativos):
    - No solo se busca eliminar el ruido.
    - Se aprende cómo suena una voz humana perfecta y frente a un audio dañado, utiliza ese conocimiento para reconstruir las componentes que faltan.

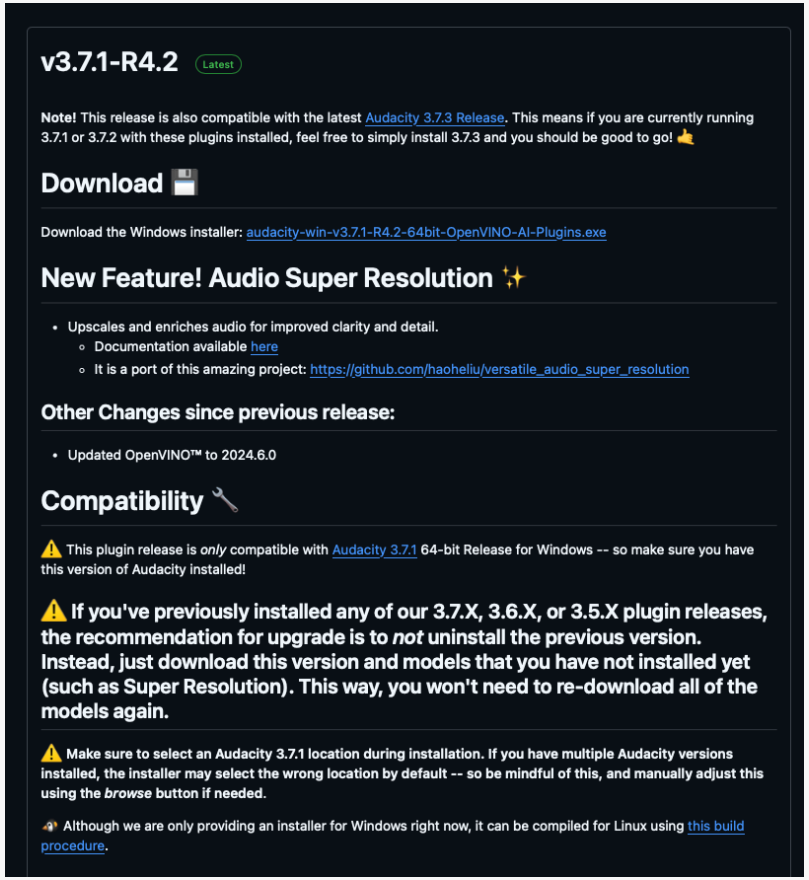
Restauración de Audios

- **Tareas:**
  - **Denoising (Eliminación de ruido):**
    - Eliminar siseos de cinta, zumbidos, tráfico o viento, dejando la voz aislada.
  - **Dereverberation (Eliminación de eco):**
    - Las grabaciones en iglesias, salas grandes o pasillos tienen un eco que hace ininteligible el discurso.
  - **Bandwidth Extension (Extensión de banda):**
    - Mejora de la calidad para tecnologías muy antiguas o audios telefónicos (por ejemplo)

Restauración de Audios

- **Herramienta Open-Source**

  - Open VINO (Plugin para Audacity):



v3.7.1-R4.2 Latest

Note! This release is also compatible with the latest Audacity 3.7.3 Release. This means if you are currently running 3.7.1 or 3.7.2 with these plugins installed, feel free to simply install 3.7.3 and you should be good to go! 🤙

Download 💾

Download the Windows installer: audacity-win-v3.7.1-R4.2-64bit-OpenVINO-AI-Plugins.exe

New Feature! Audio Super Resolution ✨

- Upscales and enriches audio for improved clarity and detail.
  - Documentation available here
  - It is a port of this amazing project: https://github.com/haoheliu/versatile_audio_super_resolution

Other Changes since previous release:

- Updated OpenVINO™ to 2024.6.0

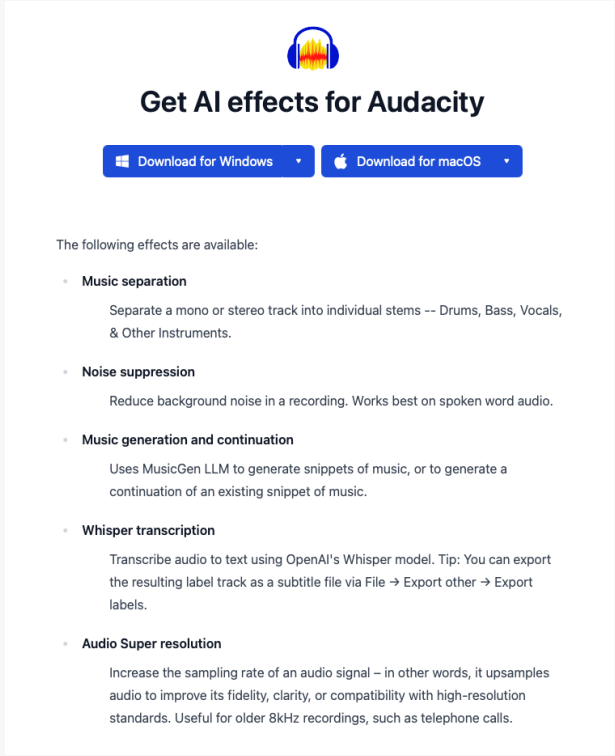
Compatibility 🔧

⚠️ This plugin release is *only* compatible with Audacity 3.7.1 64-bit Release for Windows -- so make sure you have this version of Audacity installed!

⚠️ **If you've previously installed any of our 3.7.X, 3.6.X, or 3.5.X plugin releases, the recommendation for upgrade is to *not* uninstall the previous version. Instead, just download this version and models that you have not installed yet (such as Super Resolution). This way, you won't need to re-download all of the models again.**

⚠️ **Make sure to select an Audacity 3.7.1 location during installation. If you have multiple Audacity versions installed, the installer may select the wrong location by default -- so be mindful of this, and manually adjust this using the *browse* button if needed.**

📣 Although we are only providing an installer for Windows right now, it can be compiled for Linux using this build procedure.



Get AI effects for Audacity

Download for Windows | Download for macOS

The following effects are available:

- **Music separation**

  Separate a mono or stereo track into individual stems -- Drums, Bass, Vocals, & Other Instruments.

- **Noise suppression**

  Reduce background noise in a recording. Works best on spoken word audio.

- **Music generation and continuation**

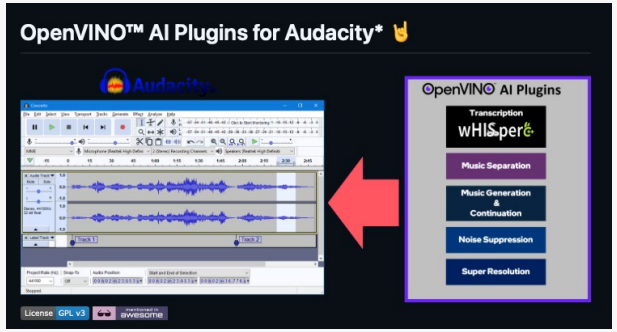
  Uses MusicGen LLM to generate snippets of music, or to generate a continuation of an existing snippet of music.

- **Whisper transcription**

  Transcribe audio to text using OpenAI's Whisper model. Tip: You can export the resulting label track as a subtitle file via File → Export other → Export labels.

- **Audio Super resolution**

  Increase the sampling rate of an audio signal – in other words, it upsamples audio to improve its fidelity, clarity, or compatibility with high-resolution standards. Useful for older 8kHz recordings, such as telephone calls.



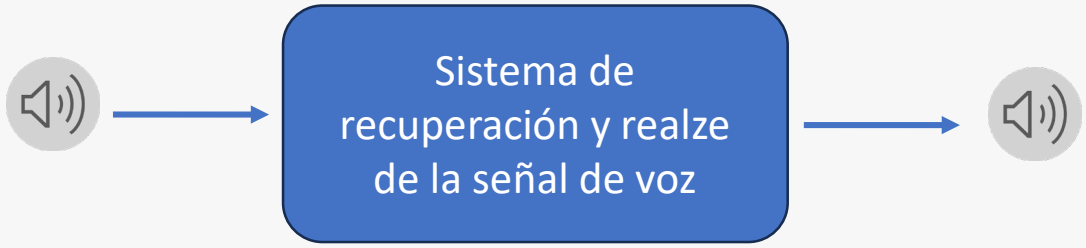
https://github.com/intel/openvino-plugins-ai-audacity

Restauración de Audios

PRUEBA DE CONCEPTO: Recuperación cintas magnetofónicas muerte de Franco

Restauración grabaciones emisión de radio en cintas magnetofónicas año 75

- Copias legales de emisión de los distintos canales de RNE
- Grabaciones con ruido, principalmente por reutilización de las cintas (grabaciones sin cabeza de borrado)
- Ancho de banda muy limitado (<1,7kHz)



Franco 4:58 https://www.rtve.es/play/audios/franco-458/

Javier Hernández Bravo

'Franco 4:58' es un podcast que trata de reconstruir cómo fue y cómo se contó la muerte de Franco en Radio Nacional de España. 50 años después de la muerte del dictador, la radio pública cuenta qué pasó la madrugada del 20 de noviembre de 1975.