Análisis y desarrollo de representaciones generales de audio

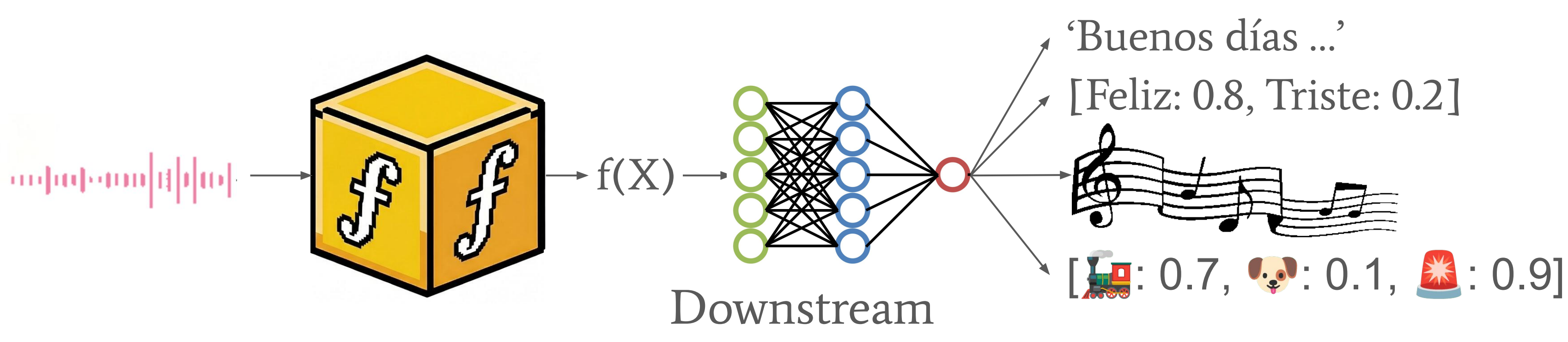
Autor: Dr. Leonardo Pepino

Directora: Dra. Luciana Ferrer

Director Adjunto: Dr. Pablo Riera

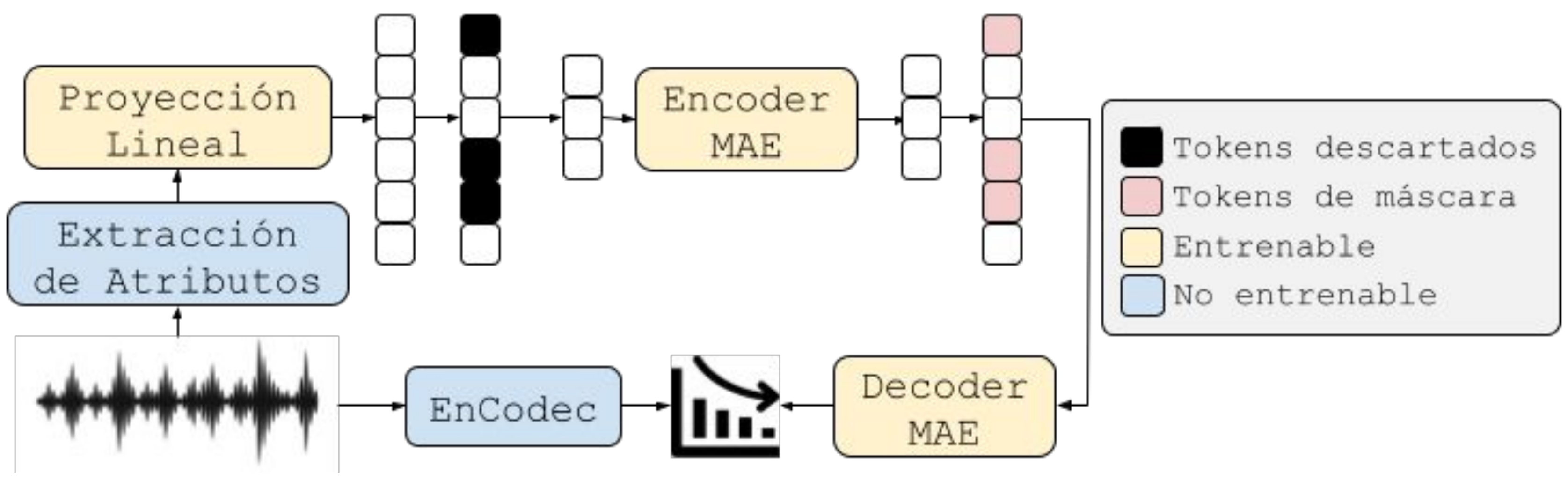
Tesis para optar por el título de Dr. en Ciencias de la Computación. Defendida el 16-12-2025

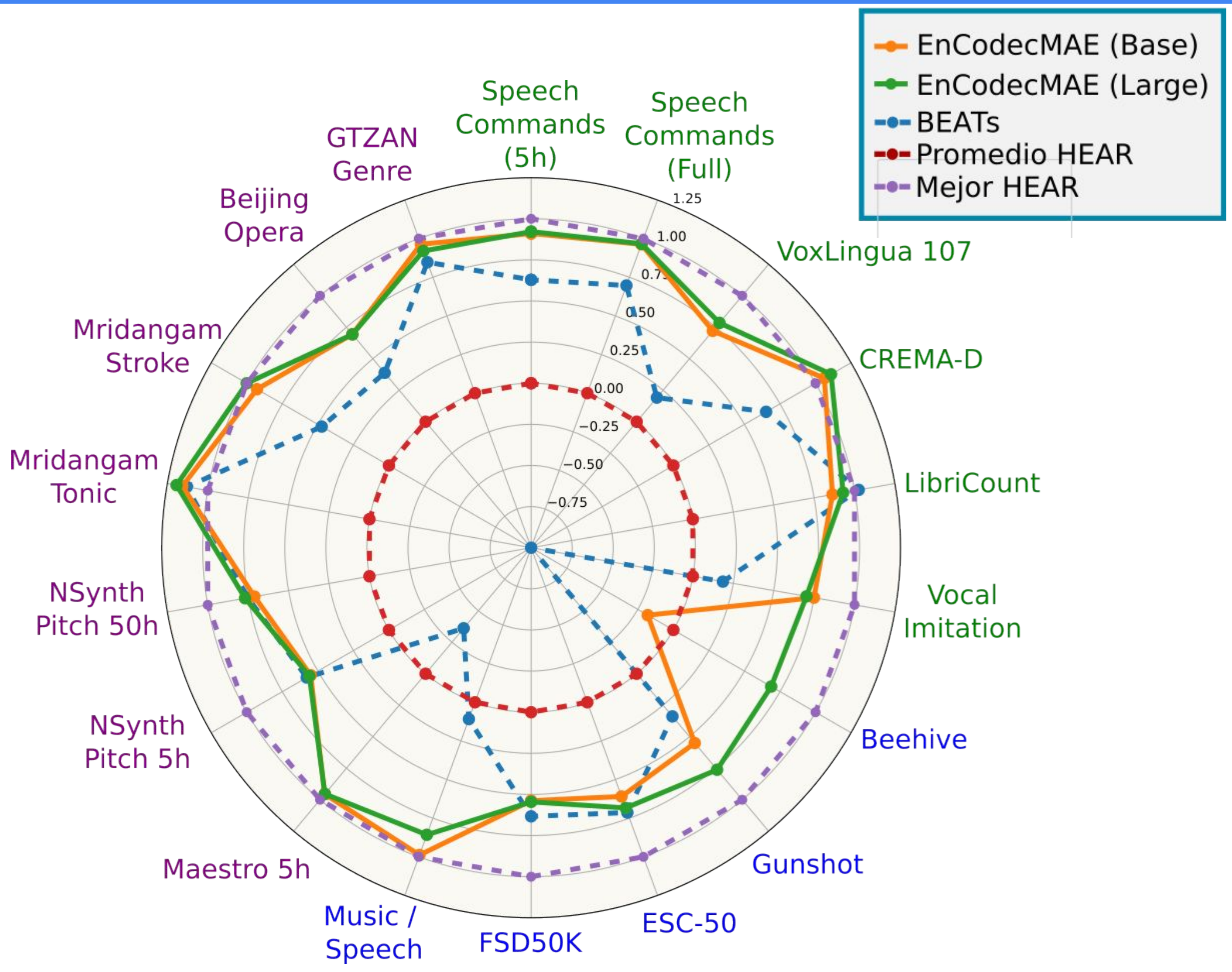
Motivación



- f es una red neuronal profunda (Upstream)
- Pre-entrenamiento eficiente y barato
- Representación general (habla, música, ambiente)
- ¿Cómo aprende a representar? ¿Como nosotros?

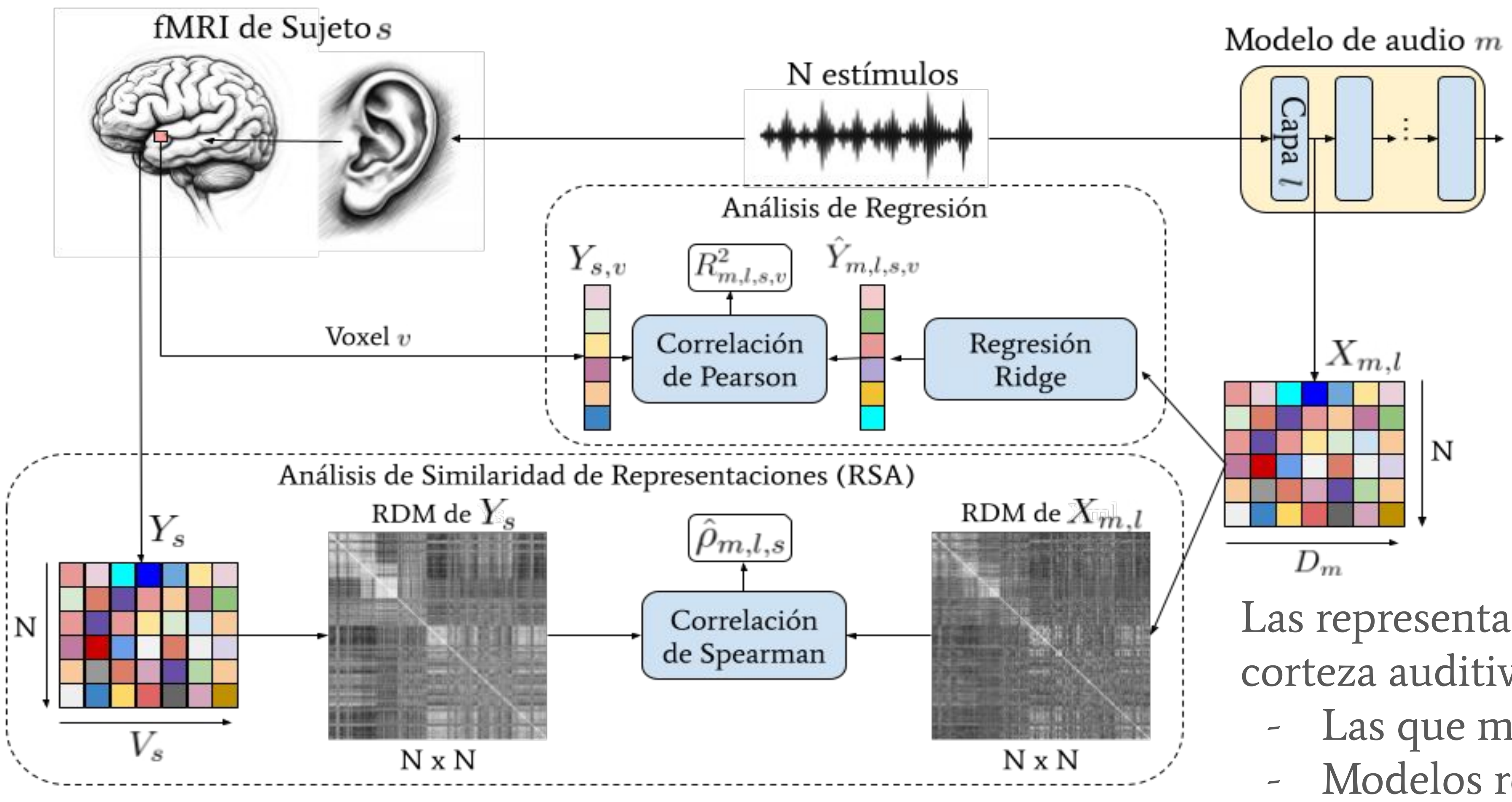
EnCodecMAE

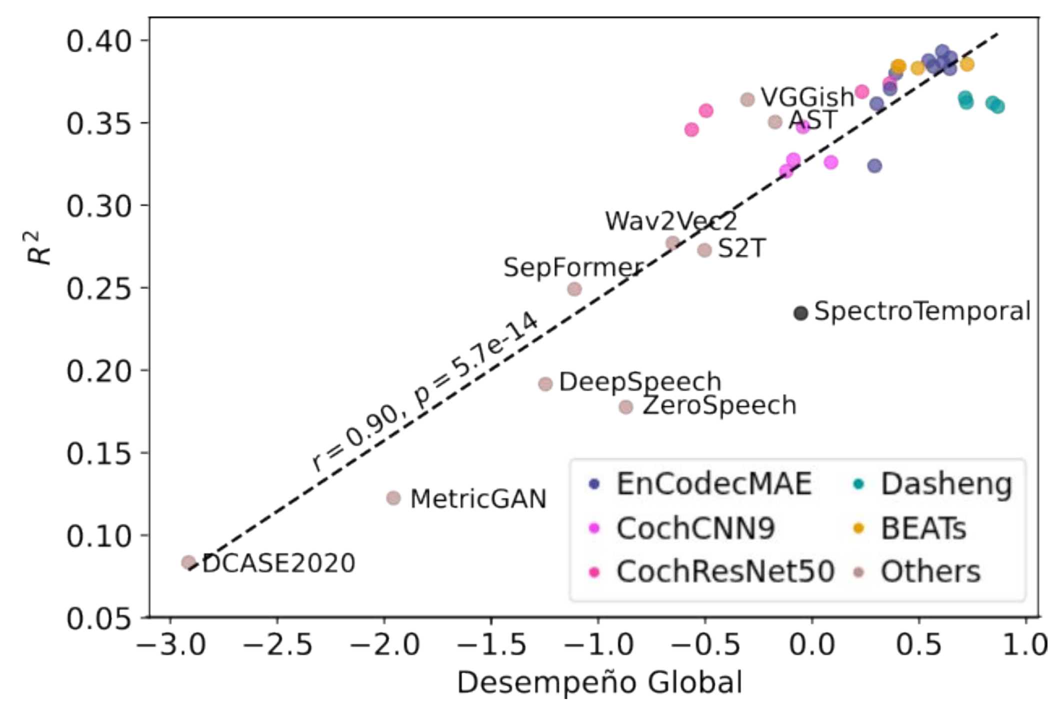




- Se entrena en 5 días con 1 o 2 RTX 3090
- 12000 horas de audio diverso: habla, música y videos de YouTube
- 75 embeddings por segundo
- Comparable o superior al estado del arte en 18 tareas de HEAREval
- Mejores resultados al aplicar una etapa de auto-entrenamiento, utilizar melspectrogramas de entrada y enmascarar el 50% de la señal
- 3 Tamaños disponibles: Small (43M), Base (86M) y Large (261M)

Representaciones de Audio vs Cerebrales





Las representaciones que mejor predicen actividad cerebral en la corteza auditiva son:

- Las que mejor se desempeñan en tareas de audición
- Modelos recientes de audio como EnCodecMAE, BEATs y Dasheng

Tanto las métricas de RSA como las de regresión incrementan durante el pre-entrenamiento de EnCodecMAE

Líneas Futuras de Investigación

Interpretabilidad: ¿Cómo logramos representaciones más interpretables?

Disentanglement: ¿Cómo hacemos para que distintas características interpretables como pitch, energía, contenido fonético, sean fácilmente accesibles en la representación?

Destilación: ¿Podemos hacer EnCodecMAEs más chicos que funcionen localmente en tiempo real en un celular?

Nuevas señales objetivo: Explorar nuevos tokenizadores de audio como señal objetivo ¿Podemos hacer un ensamble de señales objetivo especializadas en distintos dominios?

Regularización con mediciones cerebrales: ¿Se pueden aprender mejores representaciones de audio alinéandolas con mediciones cerebrales?

Representaciones bioacústicas: ¿Podemos utilizar los mismos métodos con sonidos típicos del entorno de animales? ¿Podemos entender mejor cómo procesan sonidos otras especies?