

BATS

Bridging Bridging Acoustic Transparency in Speech

Autores: Felipe Cisternas y Diego Quezada

Profesora: Raquel Pezoa

23 de Noviembre, 2023



Contenido

1. Definición del problema.
2. Marco teórico.
3. Trabajo relacionado
4. Propuesta de solución.
5. Resultados.
6. Conclusiones.

1. Definición del problema

- Modelos del estado del arte en tareas de ASR son cajas negras.
- La naturaleza física de los datos agrega una capa adicional de complejidad.

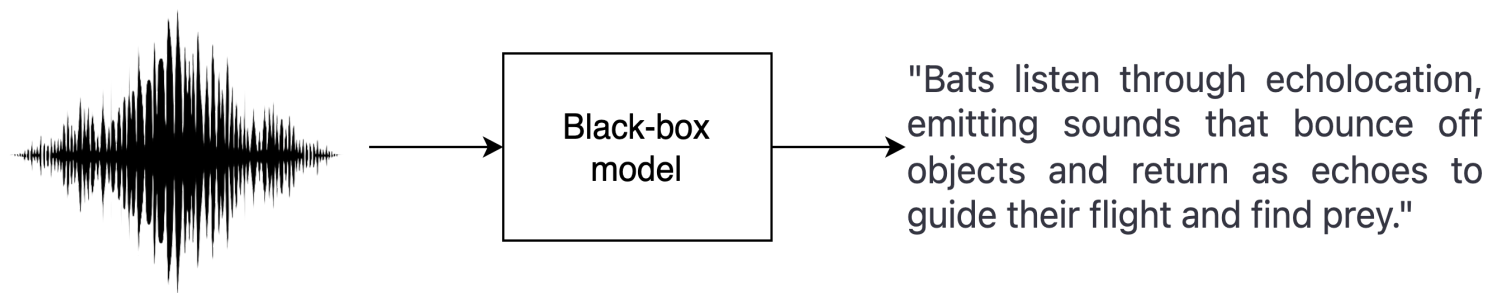


Figura 1: Esquema de un sistema de reconocimiento de voz.

1.1 Motivación

- Librerías como LIME, SHAP, Captum, etc, no permiten explicar modelos de ASR.
- **FM TEXT**: Hacer accesible la radio a personas sordas.
- ¿Cómo mejorar la transparencia y la comprensión de estos modelos?

2. Marco teórico

1. Speech recognition.
2. Representación del sonido (señal, espectrograma, y mfcc).
3. Métricas de evaluación.
4. Open ASR Leaderboard.
5. Whisper.

2.1 Speech recognition

Considerando:

- $\mathbf{X} = (x^{(1)}, x^{(2)}, \dots, x^{(T)})$: una secuencia de audio de largo T
- $\mathbf{y} = (y_1, y_2, \dots, y_N)$: una secuencia de palabras de largo N .
- P : Distribución de probabilidad condicional que relaciona \mathbf{X} con \mathbf{y} .

La tarea de reconocimiento de voz se define como:

$$f^*(\mathbf{X}) = \arg \max_{\mathbf{y}} P(\mathbf{y} | \mathbf{X} = X)$$

2.2 Representación del sonido

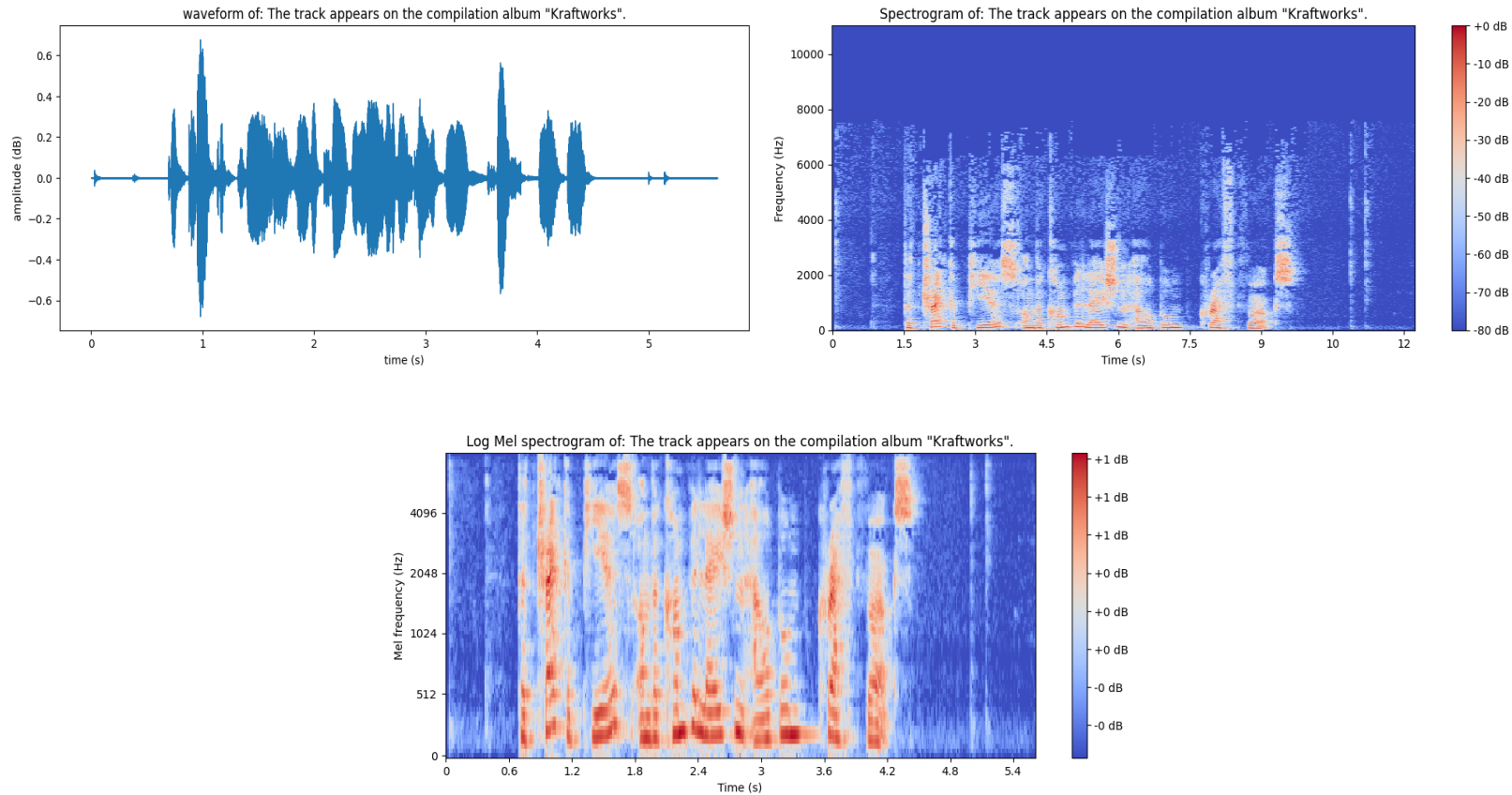


Figura 2: Distintas representaciones del sonido.
(Waveform, Spectrogram and Log Mel Spectrogram)

2.3 Métricas de evaluación

Considerando la siguiente notación:

- S : número de sustituciones.
- D : número de eliminaciones
- I : número de inserciones
- N : número de palabras en la referencia
- C : número de palabras correctas
- P : número de palabras en la predicción.

Podemos definir las siguientes métricas:

- WER (Word Error Rate) = $\frac{S+D+I}{N}$
- MER (Match Error Rate) = $\frac{S+D+I}{S+D+C}$
- WIL (Word Information Loss) = $1 - \frac{C}{N} + \frac{C}{P}$
- WIP (Word Information Preserved) = $\frac{C}{N} + \frac{C}{P}$
- CER (Character Error Rate) = $\frac{S+D+I}{N}$

CER considera P como el número de caracteres en vez de palabras.

2.4 Open ASR Leaderboard

- Competencia basada en el paper **ESB: A Benchmark For Multi-Domain End-to-End Speech Recognition (Sanchit et al. 2022)**.
- Metricas Evaluadas:
 - WER (Word Error Rate).
 - RTF (Real Time Factor).

Dataset	Domain	Speaking Style	Train (h)	Dev (h)	Test (h)	Transcriptions	License
LibriSpeech	Audiobook	Narrated	960	11	11	Normalised	CC-BY-4.0
Common Voice 9	Wikipedia	Narrated	1409	27	27	Punctuated & Cased	CC0-1.0
VoxPopuli	European Parliament	Oratory	523	5	5	Punctuated	CC0
TED-LIUM	TED talks	Oratory	454	2	3	Normalised	CC-BY-NC-ND 3.0
GigaSpeech	Audiobook, podcast, YouTube	Narrated, spontaneous	2500	12	40	Punctuated	apache-2.0
SPGISpeech	Fincancial meetings	Oratory, spontaneous	4900	100	100	Punctuated & Cased	User Agreement
Earnings-22	Fincancial meetings	Oratory, spontaneous	105	5	5	Punctuated & Cased	CC-BY-SA-4.0
AMI	Meetings	Spontaneous	78	9	9	Punctuated & Cased	CC-BY-4.0

Figura 3: Datasets ESB.

[Leaderboard](#)
[Metrics](#)
[Request a model here!](#)

model	Average WER	RTF (1e-3)	AMI	Earnings22	Gigaspeech	LS Clean	LS Other	SPGISpeech	Ted
openai/whisper-large-v3	7.7	10.3	16.01	11.3	10.02	2.03	3.91	2.95	3.9
nvidia/stt_en_fastconformer_transducer_xlarge	8.06	12.3	18.28	16.37	11.58	1.5	2.88	4.4	4.45
openai/whisper-large-v2	8.06	10.5	16.82	12.02	10.57	2.56	5.16	3.77	4.02
nvidia/stt_en_fastconformer_transducer_xlarge	8.07	14.4	18.81	16.66	11.95	1.38	2.52	4.98	4.74
distil-whisper/distil-large-v2	8.31	4.93	14.65	12.12	10.31	2.95	6.39	3.28	4.3
nvidia/stt_en_fastconformer_ctc_xlarge	8.34	5	17.62	16.44	11.61	1.69	3.4	4.91	4.64
nvidia/stt_en_conformer_ctc_large	8.39	7.5	15.97	15.83	11.59	2.06	4.16	5.6	4.42
openai/whisper-medium.en	8.5	10.7	16.43	12.59	11.13	3.02	5.84	3.41	4.15
nvidia/stt_en_fastconformer_ctc_xlarge	8.52	2.9	18.41	17.89	11.84	1.73	3.47	5.04	4.72
nvidia/stt_en_fastconformer_ctc_large	8.9	1.8	18.59	18.67	12.15	1.95	4.04	5.03	4.74
nvidia/stt_en_fastconformer_transducer_large	8.94	10.4	20.2	19.36	12.16	1.67	3.64	4.43	4.44
openai/whisper-large	9.2	10.5	17.9	15.77	11.84	3.04	6.01	3.99	4.65
nvidia/stt_en_conformer_transducer_large	9.27	21.8	22.27	19.91	12.5	1.64	3.51	4.96	5.15
distil-whisper/distil-medium.en	9.32	3.95	16.02	12.89	11.26	3.6	7.67	3.77	4.8
openai/whisper-small.en	9.34	8.3	17.88	13	11.36	3.05	7.53	3.62	4.02
nvidia/stt_en_conformer_transducer_small	10.81	17.7	20.57	18.37	13.81	2.8	6.49	6.64	6.22
openai/whisper-base.en	11.67	7.2	21.74	15.1	12.75	4.27	10.47	4.23	4.92
nvidia/stt_en_conformer_ctc_small	11.77	3.2	20.54	18.87	14.5	3.58	7.89	7.75	7.15
patrickvonplaten/wav2vec2-large-960h-lv60-self-4-gram	13.65	20.1	28.85	23.06	16.13	1.75	3.55	9.22	7.25
facebook/wav2vec2-large-960h-lv60-self	14.47	2.5	30.84	25.04	16.73	1.73	3.74	9.53	7.35
openai/whisper-tiny.en	14.96	9.1	24.68	19.35	14.08	5.66	15.38	5.82	5.96

Figura 4: Open ASR Leaderboard, Fuente: Huggingface

2.5 Whisper

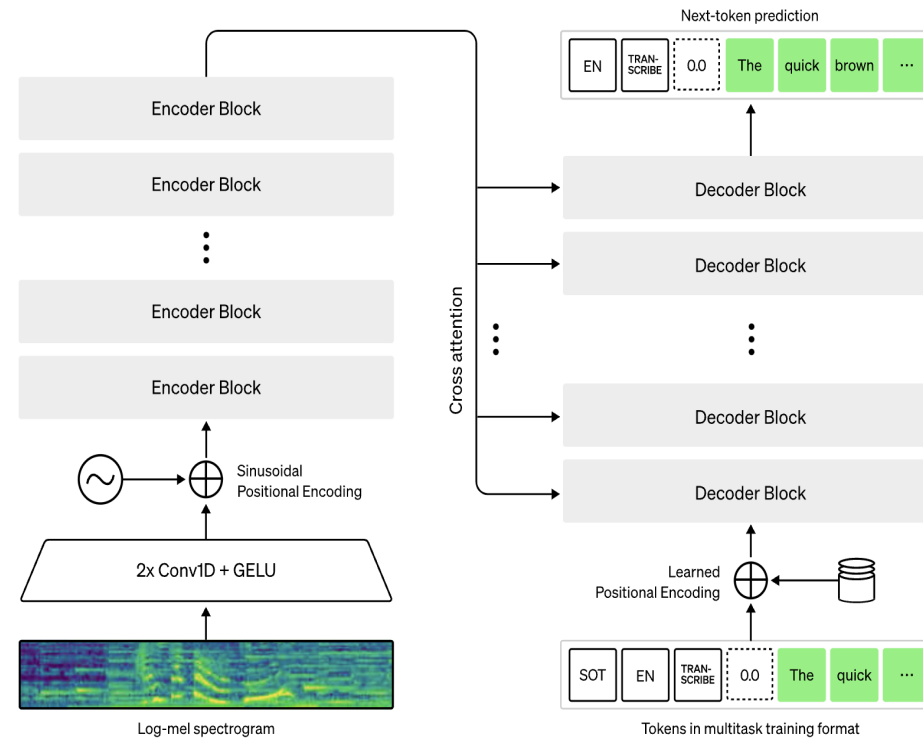


Figura 5: Arquitectura Whisper, Fuente: OpenAI

3. Trabajo relacionado

Se han propuesto explicaciones para distintas tareas:

- **Reconocimiento de voz:** Segmentos de audio que son causas mínimas y suficientes.
- **Reconocimiento de fonemas:** Importancia de segmentos de audio.
- **Etiquetado de música:** Importancia de fuentes de audio.

Publicación	Tarea	Métodos
X. Wu, et al. (2020)	Reconocimiento de voz	SFL, Causal, LIME (*)
Haunschmid, et al. (2020)	Etiquetado de música	LIME (*)
X. Wu, et al. (2023)	Reconocimiento de fonemas	LIME (*)

Tabla 1: Resumen de trabajos relacionados

(*) Versión modificada de LIME.

4. Propuesta de solución

El entorno de experimentación se define a continuación:

- Conjunto de datos: Common Voice 11.
- Modelo: Whisper versión Tiny.
- CPU: Apple M1 Pro 10 Cores.
- GPU: Apple GPU 16 Cores.
- RAM: 16GB LPDDR5.
- OS: macOS 14.0.
- Software: Python 3.10, PyTorch 2.1.0, 4.3.5.0.

4.1 SLIME

- Representación: Vector booleano para ausencia o presencia de un segmento.
- Vecindad: Muestra aleatoria de una distribución binomial con probabilidad $p = 0.5$.
- Modelo interpretable: Regresión lineal o árbol de decisión.
- Tarea: Predicción Distancia de Levenshtein respecto a la transcripción original.
- Explicación: Coeficientes asociados a cada segmento.

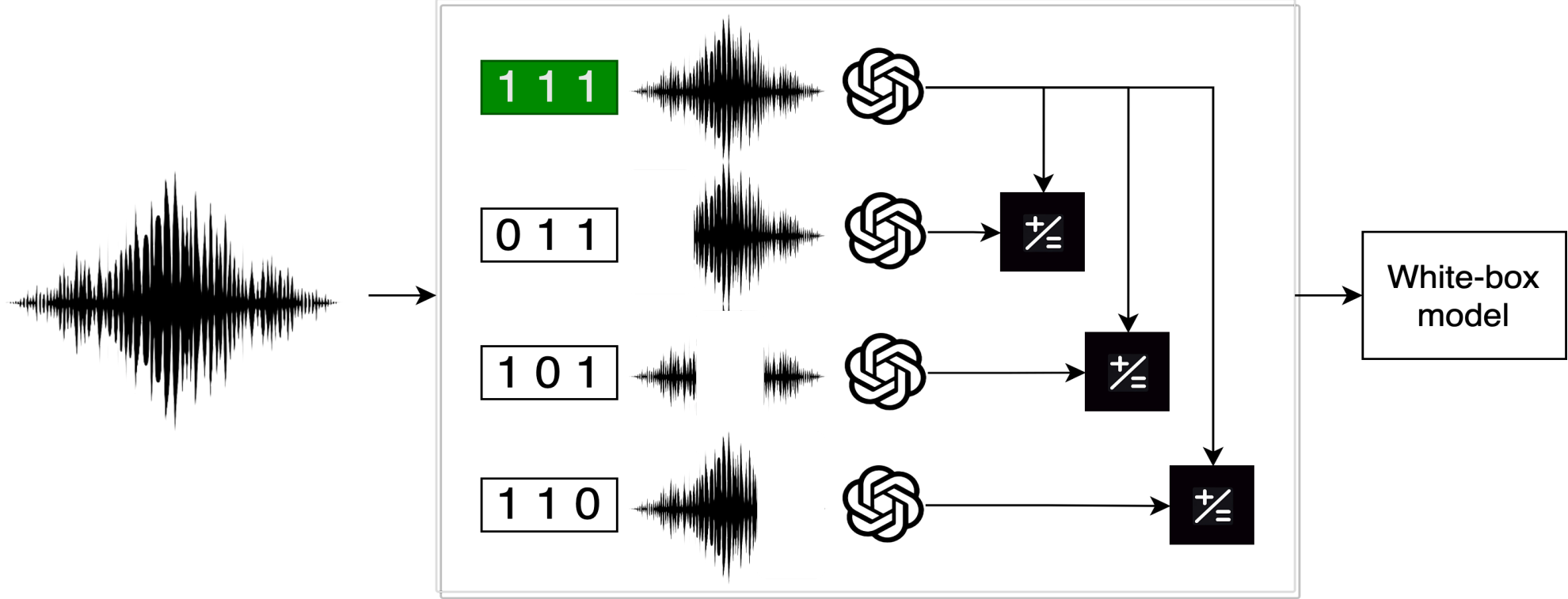


Figura 6: SLIME.

4.2 Borrado de representaciones

- Representación: Espectrograma de MEL (80,3000).
- Calculo de importancia: WER, MER, WIL, WIP, CER
- Comparación entre espectrograma original y espectrograma con dimensiones borradas.
- $$I(d) = \frac{1}{|E|} \sum_{x \in E} \frac{S_M(x,y) - S_M(x,y, \neg d)}{S_M(x,y)}$$
- Explotación: Bandas de frecuencia más importantes.

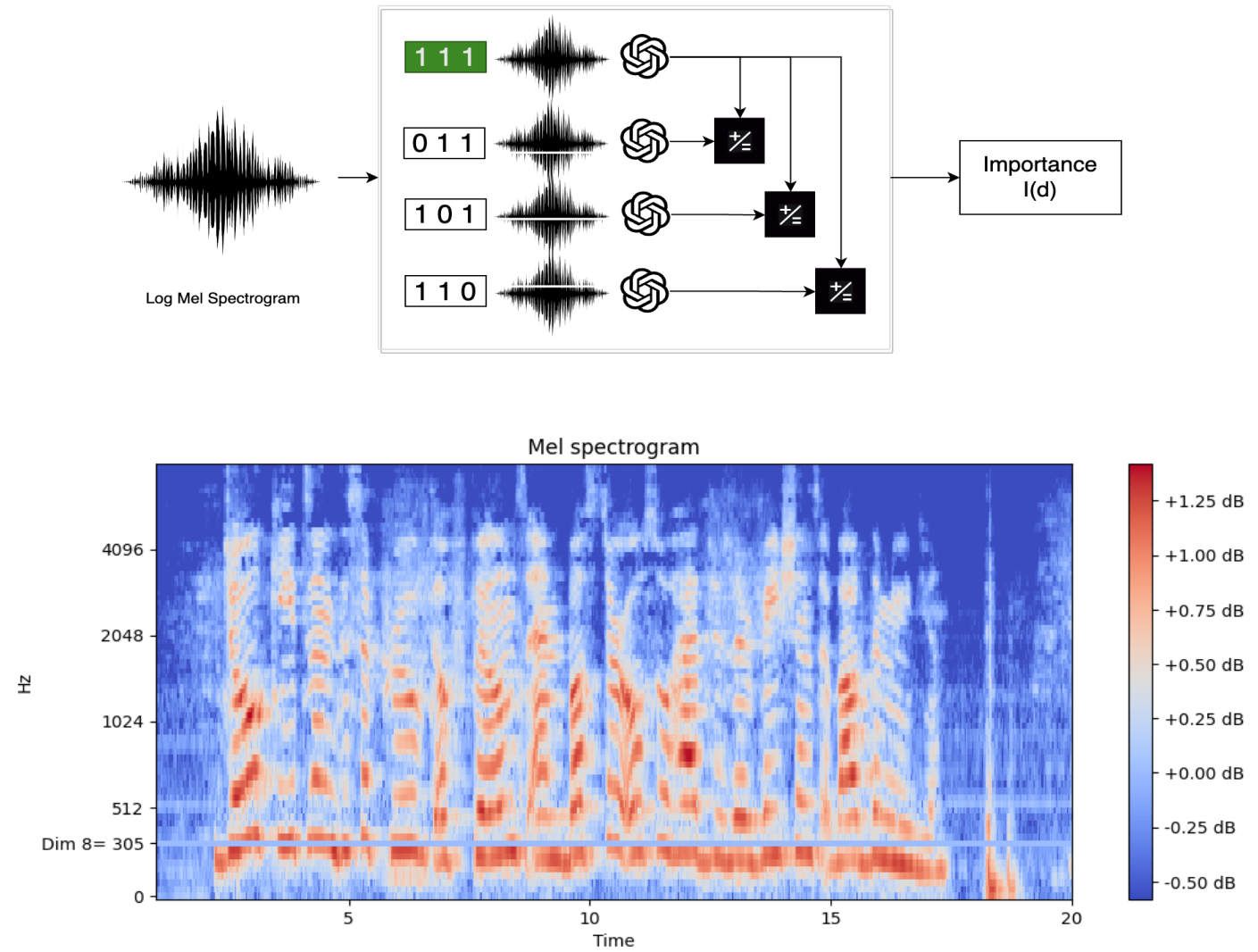


Figura 7: Borrado de Representaciones y Ejemplo con un Espectrograma Real.

5. Resultados

5.1 SLIME

- y : it is a busy market town that serves a large surrounding area.
- y' : it is a busy market town that serves a large **surrounded** area.
- A continuación, se explica la transcripción de Whisper usando SLIME.

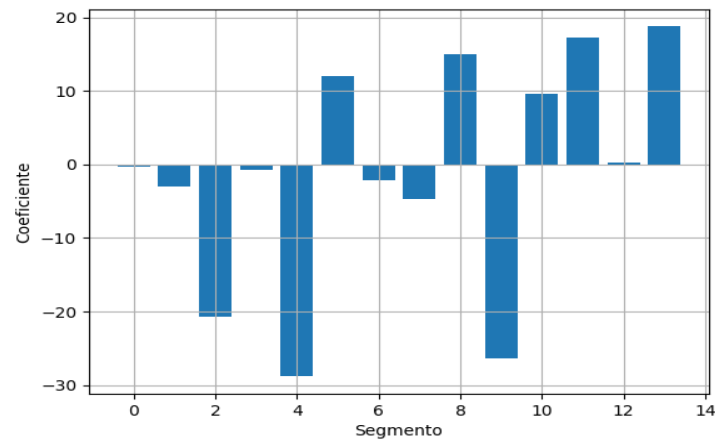


Figura 8: Coeficientes de g

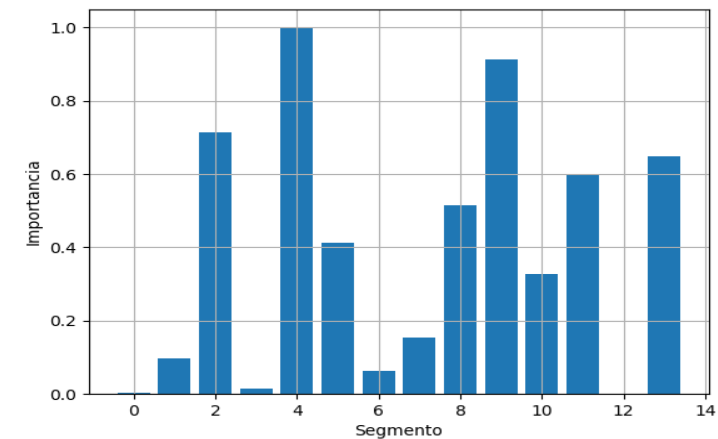


Figura 9: Importancias SLIME

5.2 Representation Erasure

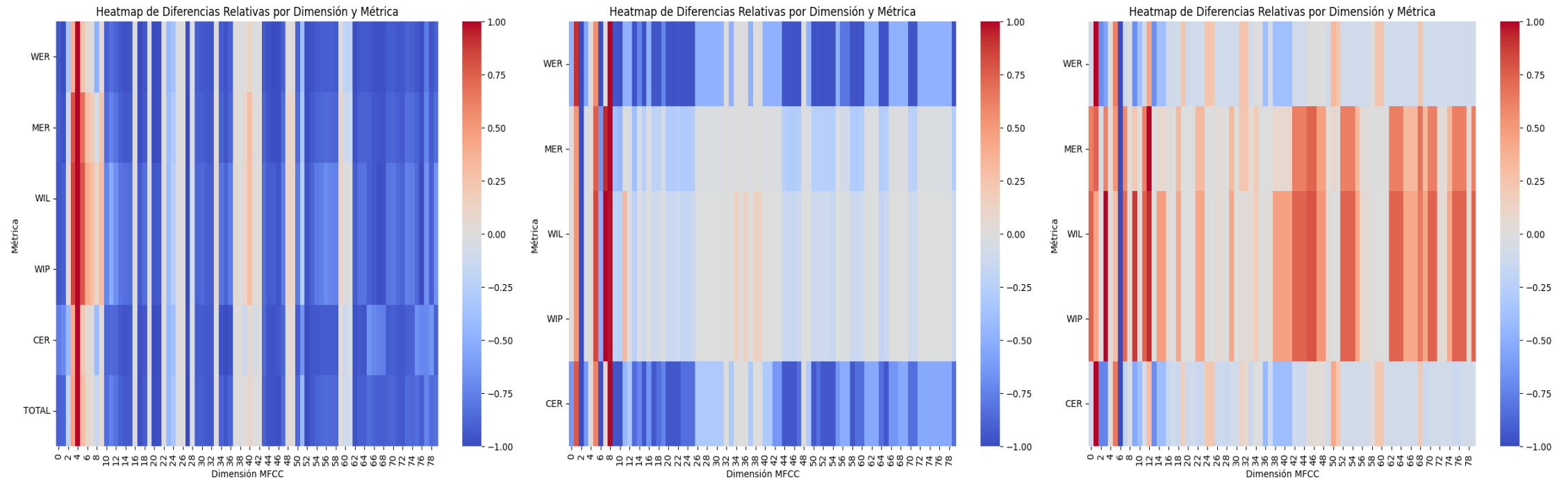


Figura 10: Importancia de las dimensiones del espectrograma

En la primera visualización se consideran todas las dimensiones.

En la segunda se elimina la dimensión 4 y en la tercera la dimensión 4 y 8.

6. Conclusiones

- Se estudió la explicabilidad de Whisper: un modelo del estado del arte en reconocimiento de voz.
- Se propuso SLIME: una adaptación de LIME para el reconocimiento de voz inspirada en LIME-TS (X. WU, et al. 2023).
- SLIME provee explicaciones escuchables.
- SLIME destaca la importancia de cada segmento de audio.
- *Representation erasure* provee explicaciones visuales.
- *Representation erasure* evidencia la importancia de las frecuencias bajas.

6.1 Trabajo futuro

Como trabajo futuro se propone:

- BATS: Librería para explicar modelos de ASR.
- Aprendizaje reforzado para aprender la mejor representación de los datos.

Muchas gracias por su atención.