# A Simple Method to Enhance Pre-trained Language Models with Speech Tokens for Classification

Nicolas Calbucura **José Guillen** Valentin Barriere

*En proceso de revisión en EMNLP 2026*
*Preprint: arXiv:2512.07571*

3 de Junio, 2026

# Contenido

# Introducción

# Comunicación Humana

La comunicación humana es multimodal [Fröhlich et al., 2019].

- Se compone de:
  - Lenguaje hablado
  - Lenguaje escrito
  - Lenguaje no verbal
    - Gestos
    - Expresiones faciales
    - Postura corporal
- Puede tener fines informativos, persuasivos, entre otros.

# Especificaciones del audio

- El habla tiene un **número variable** de unidades por secuencia.
- Es una **secuencia larga** y **sin fronteras** de segmento predefinidas.
- Es **continua**, sin un diccionario de unidades predefinido (a diferencia del texto).
- Algunas tareas requieren **información distinta** (p. ej. ASR vs. identificación de hablante).

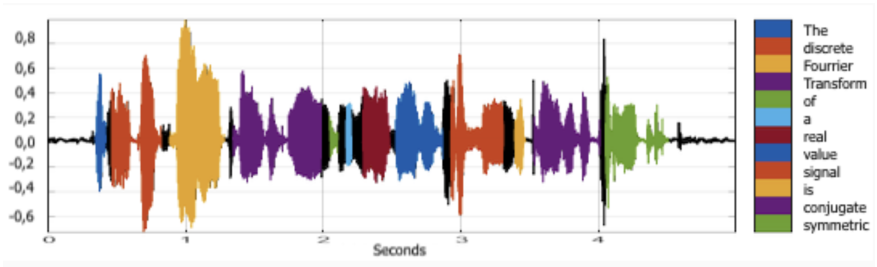
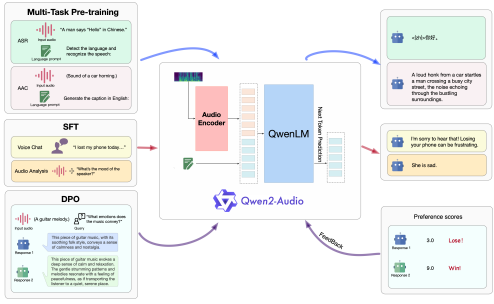



Figure: El habla es continua, mientras que el texto es discreto.

# El reto de integrar audio

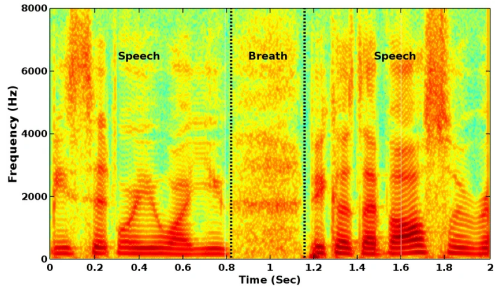
- Los métodos que integran audio (SpeechVerse, Qwen2-Audio, Audio-Flamingo) se basan en **varias etapas de pre-entrenamiento** (ASR + *instruction tuning*).




SpeechLLM: pre-entrenamiento (ASR) + *instruction tuning*.

# El reto de integrar audio

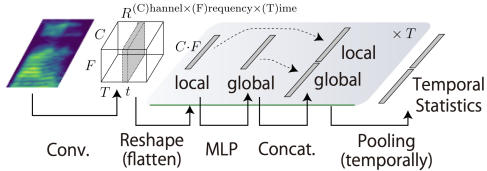
- Los métodos que integran audio (SpeechVerse, Qwen2-Audio, Audio-Flamingo) se basan en **varias etapas de pre-entrenamiento** (ASR + *instruction tuning*).
- Las señales de audio producen representaciones latentes **mucho menos compactas** que el lenguaje natural: ≈**50 ó 12 Hz** frente a ≈**2 Hz**.




Alta resolución temporal → muchos más tokens que el texto.

# El reto de integrar audio

- Los métodos que integran audio (SpeechVerse, Qwen2-Audio, Audio-Flamingo) se basan en **varias etapas de pre-entrenamiento** (ASR + *instruction tuning*).
- Las señales de audio producen representaciones latentes **mucho menos compactas** que el lenguaje natural: ≈**50 ó 12 Hz** frente a ≈**2 Hz**.
- Si el audio se colapsa en una **representación única** (p. ej. BYOL-A), se reduce la longitud pero se **pierde la secuencialidad** del audio al fusionarlo.




Representación única: el audio se resume en un solo vector → se pierde la secuencia.

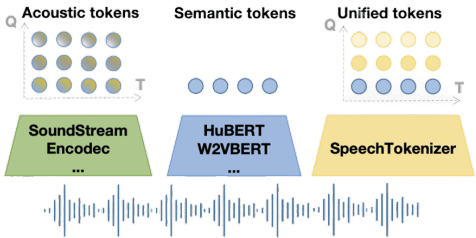
**¿Cómo aprovechar un LLM ya entrenado para integrar una nueva modalidad acústica, considerando sus problemáticas?**

# Tokenización del habla y propuesta

- **Speech Tokenization**: convertir señales de audio continuas en **tokens discretos** que capturan información **semántica y/o acústica**, procesables por modelos diseñados para texto.
- Dos categorías: tokens **semánticos** (p. ej. HuBERT) y tokens **acústicos** (p. ej. EnCodec), obtenidos por separado.

**Nuestra propuesta**

**En vez de comprimir la secuencia**, el método **identifica los tokens acústicos específicos** que cargan más información para la tarea.




Audio continuo → tokens discretos (semánticos + acústicos).

# ¿Qué tareas abordaremos?

# Falacias

**Falacias Argumentativas**

Argumentos inválidos o fallos lógicos que pueden debilitar la validez de un argumento.

Ejemplos:

- *Ad Hominem*: Atacar a la persona en lugar del argumento.
- *Apelación a la autoridad*: Afirmar que algo es verdadero porque una autoridad lo dice.

# Falacias

## AFD: Detección de Falacias

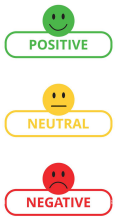
Tarea en la cual se identifica si una frase contiene una falacia o no. Dos clases: "falacia" y "no falacia".

## AFC: Clasificación de Falacias

Tarea que consiste en categorizar falacias en argumentos. Múltiples clases, cada una representando un tipo específico de falacia.

# Sentimiento y Emoción

## Sentimiento




Ejemplos:

- Una crítica de cine entusiasta → *positivo*.
- Una queja sobre un producto → *negativo*.

## Emoción


Ejemplos:

- Reconocer si el hablante expresa *tristeza* o *sorpresa*.
- Una misma frase puede mezclar varias emociones.

# Sentimiento y Emoción: Tareas de clasificación

**SENT: Análisis de Sentimiento**

Tarea en la cual se clasifica la polaridad de una frase. Dos clases: "positivo" y "negativo".

**EMO: Clasificación de Emoción**

Tarea multi-etiqueta sobre seis emociones: alegría, tristeza, enojo, miedo, asco y sorpresa. Una frase puede tener varias a la vez.

# Objetivos

## Objetivos del Trabajo

1. **Evaluar modelos "básicos" de Machine Learning** (RoBERTa + codificadores de audio) en las tareas de clasificación, usando representaciones textuales y auditivas.

# Objetivos

## Objetivos del Trabajo

1. **Evaluar modelos "básicos" de Machine Learning** (RoBERTa + codificadores de audio) en las tareas de clasificación, usando representaciones textuales y auditivas.
2. **Evaluar el rendimiento de LLMs** en las mismas tareas, en modo unimodal y frente a SpeechLMs.

# Objetivos

## Objetivos del Trabajo

1. **Evaluar modelos "básicos" de Machine Learning** (RoBERTa + codificadores de audio) en las tareas de clasificación, usando representaciones textuales y auditivas.
2. **Evaluar el rendimiento de LLMs** en las mismas tareas, en modo unimodal y frente a SpeechLMs.
3. **Proponer e integrar tokens de audio en un LLM** mediante selección de tokens relevantes (lasso) y pre-entrenamiento de sus embeddings, evaluando su aporte sobre el modelo unimodal.

# Objetivos

## Objetivos del Trabajo

1. **Evaluar modelos "básicos" de Machine Learning** (RoBERTa + codificadores de audio) en las tareas de clasificación, usando representaciones textuales y auditivas.
2. **Evaluar el rendimiento de LLMs** en las mismas tareas, en modo unimodal y frente a SpeechLMs.
3. **Proponer e integrar tokens de audio en un LLM** mediante selección de tokens relevantes (lasso) y pre-entrenamiento de sus embeddings, evaluando su aporte sobre el modelo unimodal.
4. **Analizar el método**: tipo de tokens (acústicos vs. semánticos), forma de selección (lasso vs. aleatoria) y tamaño del vocabulario.

# Conceptos y Materiales

# Representaciones textuales y auditivas


Figure: Representación Bag of Words.

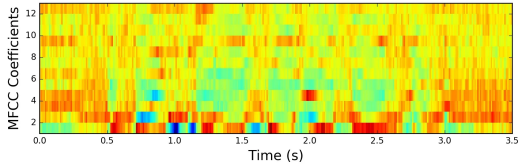



Figure: Representación de coeficientes cepstrales en frecuencia Mel (MFCCs).

# Representaciones textuales y auditivas: BYOL-A

- **BYOL-A** [Niizumi et al., 2023]: método **auto-supervisado** para aprender representaciones de audio de propósito general.
- Transforma un segmento de audio (∼1-10 s) en **un único vector**, proyectado al espacio del LLM con una capa lineal.
- Conexión con la motivación: al colapsar el audio en un solo vector, se **pierde la secuencialidad** del audio al fusionarlo.

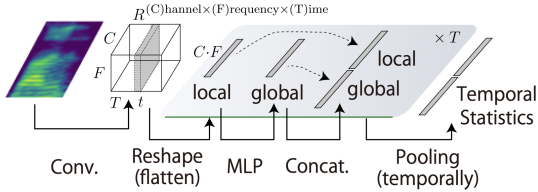



Figure: BYOL-A: el audio se resume en un único *embedding*.

# Representaciones textuales y auditivas: tokens de LLM

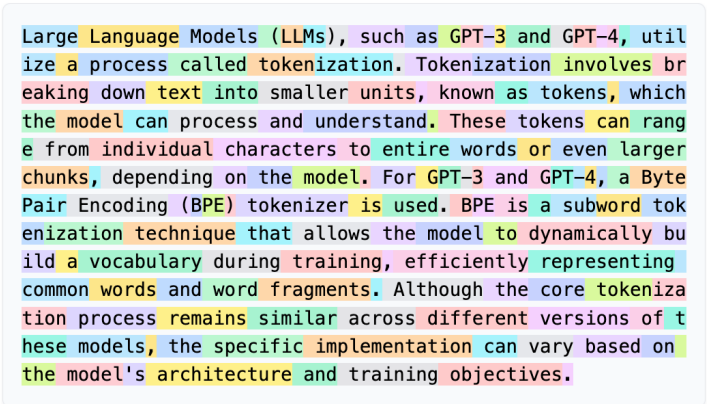



Figure: Descomposición de texto en tokens por un LLM.

# Representaciones textuales y auditivas: embeddings

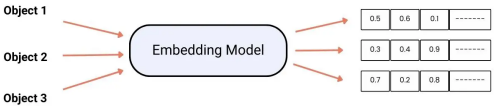



Figure: Transformación de tokens en embeddings

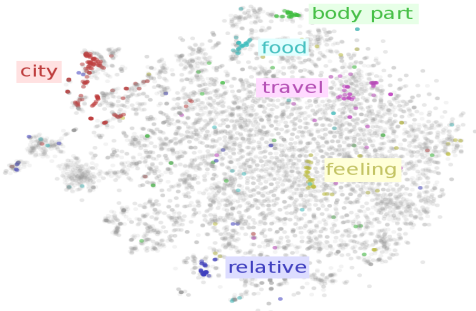



Figure: Visualización de embeddings en 2D

# SpeechLLMs: Qwen2-Audio

- Los **SpeechLLMs** (p. ej. Qwen2-Audio [Chu et al., 2024]) integran audio dentro de un LLM.
- Entrenamiento en **varias etapas**: pre-entrenamiento masivo (a menudo con **ASR**) + *instruction tuning* sobre muchas tareas.
- Requieren **enormes cantidades de datos** de pre-entrenamiento para buen rendimiento general [Thimonier et al., 2025].

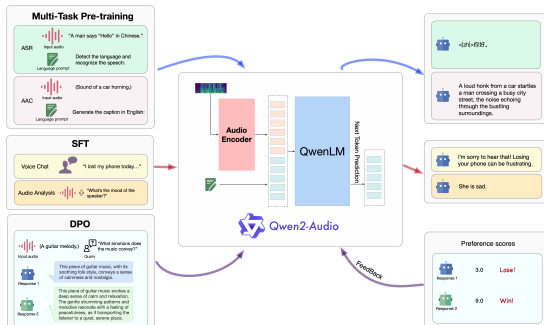



Figure: SpeechLLM: pre-entrenamiento (ASR) + *instruction tuning*.

# Nuestra propuesta

# SpeechTokenizer: Tokenizador del habla

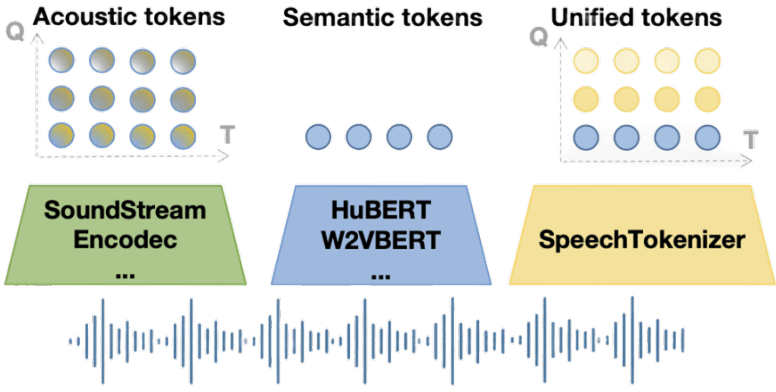



Figure: Diferencia entre otros tokenizadores y SpeechTokenizer.

# Datasets: Falacias

- **MMUSED-fallacy**: dataset para AFD y AFC.
  - 18,910 ejemplos en total, de los cuales 1,457 son falacias.
  - 6 clases distintas de falacias. Clases desbalanceadas.
  - Los ejemplos tienen una duración promedio de 10.40 segundos.
- **MMUSED**: dataset del cual se origina MMUSED-fallacy. Se utiliza para entrenar los embeddings de audio.

- Datasets utilizados en Shared Task 12th Workshop on Argument Mining at ACL 2025.
- **Ejemplos**

| Snippet | Fallacy Category |
|---|---|
| *the same kind of woolly thinking* | Appeal to Emotion |
| *As George Will said the other day, "Freedom on the march; not in Russia right now."* | Appeal to Authority |
| *Governor Carter apparently doesn't know the facts.* | Ad Hominem |
| *We won the Cold War because we invested and we went forward.* | False Cause |
| *And if we don't act today, the problem will be valued in the trillions.* | Slippery Slope |
| *We have to practice what we preach.* | Slogan |

# Falacias: ejemplo real

**Ejemplo (debate presidencial)**

*"this isn't just a Southern problem; it's a Northern problem and a Western problem; it's a problem for all of us.."*

- Falacia: **Appeal to Emotion (AE)**[a]

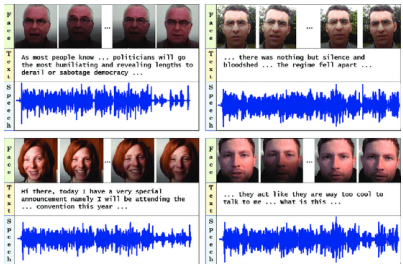
[a] En el contexto del debate (una guerra), el orador enmarca la amenaza como un problema que nos afecta *a todos*, buscando provocar **miedo** en el público —una de las seis emociones básicas de Ekman— en lugar de ofrecer evidencia o un argumento lógico.


Clip de MMUSED-fallacy — youtu.be/2gKpYUpd4PE

# Datasets: Sentimiento y Emoción

- **CMU-MOSEI**: dataset para SENT y EMO.
  - Más de 23,500 enunciados de más de 1,000 videos de YouTube.
  - **Sentimiento**: clasificación binaria de la polaridad (positivo/negativo).
  - **Emoción**: multi-etiqueta sobre 6 emociones (alegría, tristeza, enojo, miedo, asco, sorpresa).

- **Ejemplos**



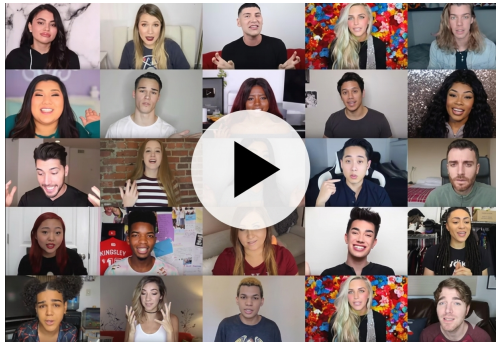
# Sentimiento y Emoción: ejemplo real

## Ejemplo (CMU-MOSEI)

*"Do you need to wait to be admitted to apply for financial aid?"*

- Solo texto → negativo / neutro
- Texto + audio → positivo / happy[a]

[a] En texto, la pregunta suena cruda o neutra; pero en el clip es un anuncio (tipo marketing) dicho en un tono **muy alegre**. El audio aporta ese afecto positivo que el texto por sí solo no capta.



Clip de CMU-MOSEI — youtu.be/QvjJg_GjcvE

# Métodos

# Método principal: Paso 1

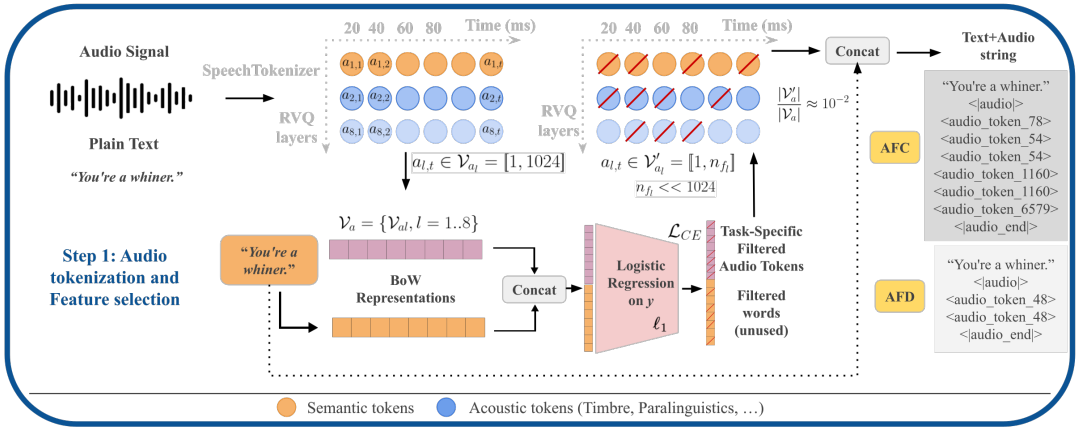



Figure: Paso 1 del método principal. Selección de tokens de audio relevantes para la tarea.

# Método principal: Paso 1

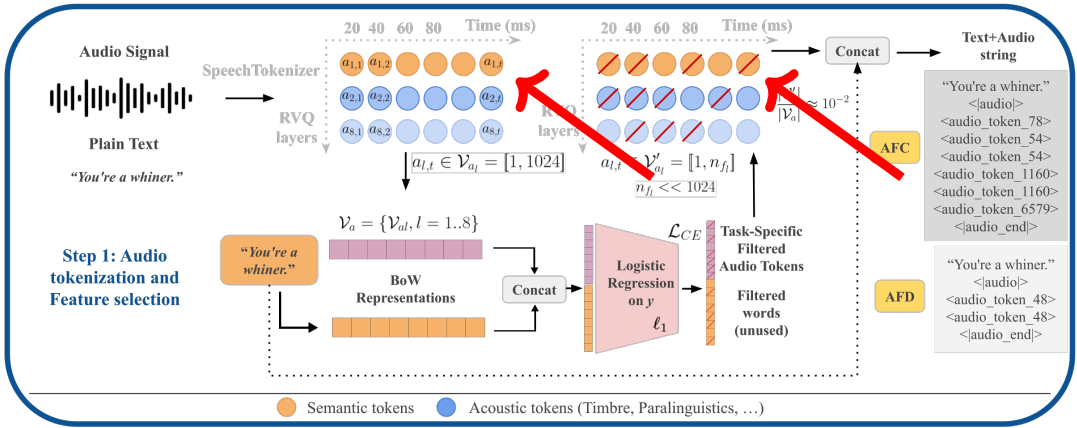



Figure: Paso 1 del método principal. Selección de tokens de audio relevantes para la tarea.

# Método principal: Paso 1

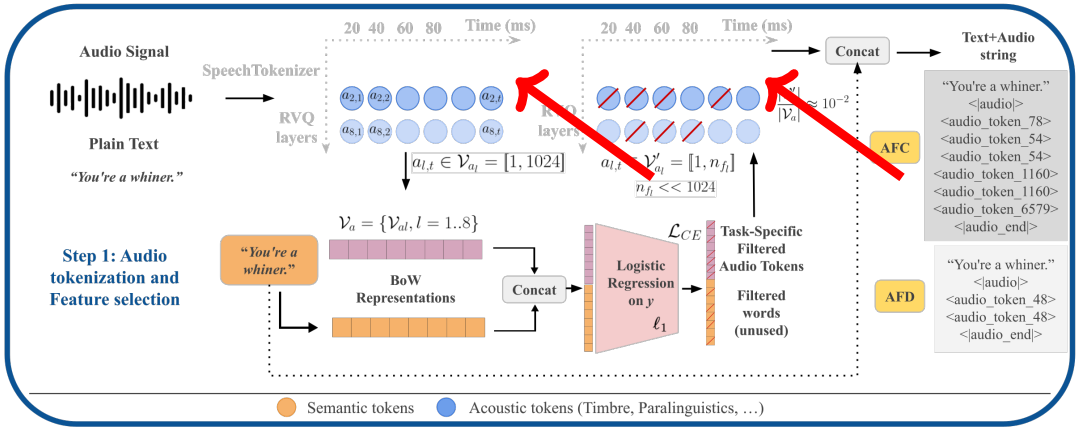



Figure: Paso 1 del método principal. Selección de tokens de audio relevantes para la tarea.

# Método principal: Paso 1

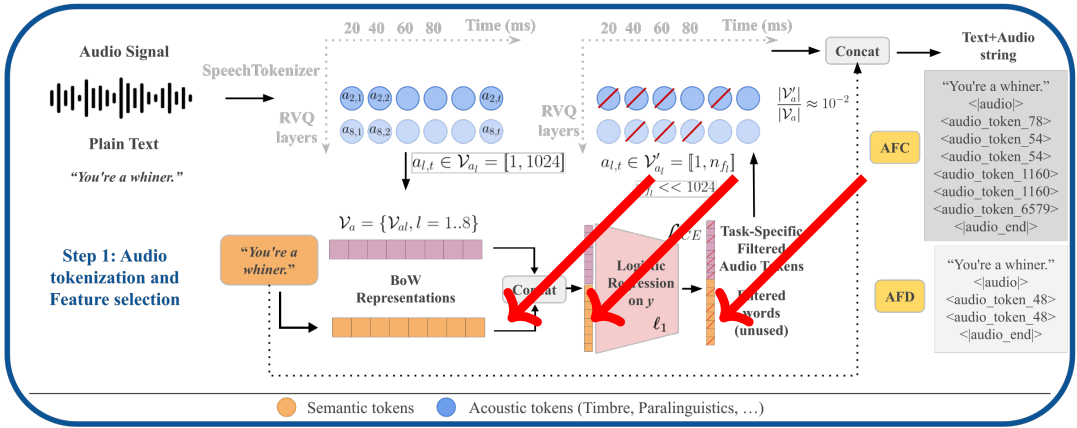



Figure: Paso 1 del método principal. Selección de tokens de audio relevantes para la tarea.

# Método principal: Paso 1

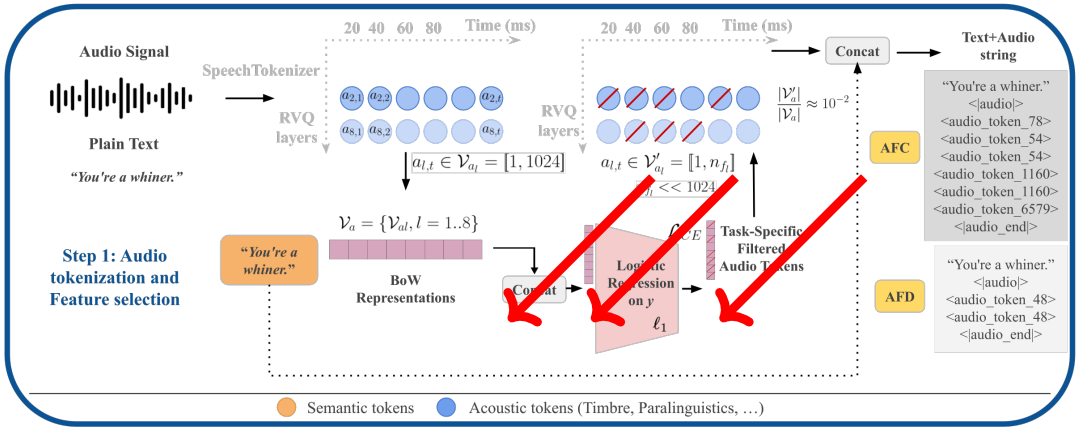



Figure: Paso 1 del método principal. Selección de tokens de audio relevantes para la tarea.

# Método principal: Paso 2

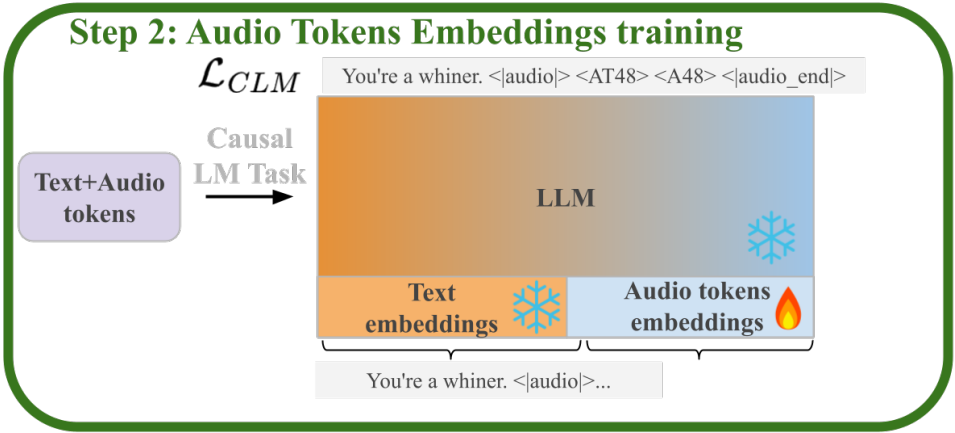



Figure: Paso 2 del método principal. Pre-entrenamiento de los embeddings de audio.

# Método principal: Paso 3

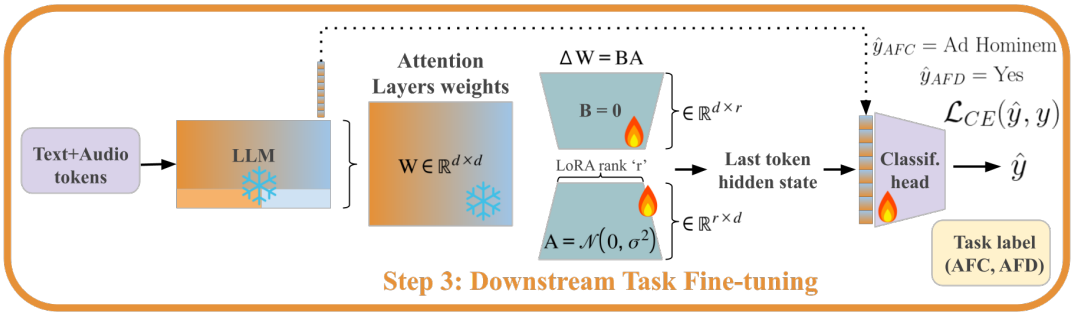



Figure: Paso 3 del método principal. Fine-tuning del LLM para tareas específicas.

# Metodología: paradigmas de entrenamiento y evaluación

- **SFT** (Supervised Fine-Tuning): se **ajustan los parámetros** del modelo con ejemplos **etiquetados** de la tarea *(p. ej. ROBERTA)*.
- **ICL** (In-Context Learning): el modelo resuelve la tarea a partir de unos pocos ejemplos en el **prompt**, **sin** actualizar sus parámetros *(Qwen2-Audio-7B)*.
- **LoRA**: fine-tuning **eficiente en parámetros**; se entrenan adaptadores de bajo rango en lugar de todo el modelo.
- **Seeds**: cada experimento se repite con varias **semillas** aleatorias (5) y se promedia, para resultados robustos.
- **Bagging**: se combinan las predicciones de las distintas semillas (*probability-based*) para mayor robustez y precisión.

# Baselines

- BoW → Texto
- MFCCs → Audio
- BoW + MFCCs → Texto + Audio
- RoBERTa → Texto
- LSTM + WavLM → Audio
- RoBERTa + WavLM → Texto + Audio
- Llama3-3B + BYOL-A → Texto + Audio
- Qwen2-Audio-7B → Texto + Audio
- Se utilizaron varias seeds para los experimentos. También se utilizó bagging en algunos casos.

# Resultados

# Resultados para AFC

| Task | Method | Bagging | Text | Audio | Text + Audio |
|---|---|---|---|---|---|
| AFC | RoBERTa + Whisper† [Tahir et al., 2025] | ✗ | 0.486 | 0.159 | 0.461 |
| | RoBERTa + HuBERT† [Pittiglio, 2025] | | 0.444 | **0.356** | 0.440 |
| | (MM-)RoBERTa + WavLM† | | 0.393 | 0.064 | 0.382 |
| | Qwen2-Audio-7B (ICL) [Chu et al., 2024] | ✗ | – | – | 0.170 |
| | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | | – | – | 0.507 |
| | Llama3-3B+BYOL-A (SFT) | | 0.523 | – | – |
| | Our method | | – | – | 0.548 |
| | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | ✓ | – | – | 0.515 |
| | Llama3-3B+BYOL-A (SFT) | | **0.537** | – | – |
| | Our method | | – | – | **0.592** |

Table: Resultados obtenidos para la tarea AFC, considerando los mejores resultados de la Shared Task.

# Resultados para AFC

| Task | Method | Bagging | Text | Audio | Text + Audio |
|---|---|---|---|---|---|
| AFC | RoBERTa + Whisper† [Tahir et al., 2025] | ✗ | 0.486 | 0.159 | 0.461 |
| | RoBERTa + HuBERT† [Pittiglio, 2025] | | 0.444 | **0.356** | 0.440 |
| | (MM-)RoBERTa + WavLM† | | 0.393 | 0.064 | 0.382 |
| | Qwen2-Audio-7B (ICL) [Chu et al., 2024] | ✗ | – | – | 0.170 |
| | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | | – | – | 0.507 |
| | Llama3-3B+BYOL-A (SFT) | | 0.523 | – | – |
| | Our method | | – | – | 0.548 |
| | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | ✓ | – | – | 0.515 |
| | Llama3-3B+BYOL-A (SFT) | | **0.537** | – | – |
| | Our method | | – | – | **0.592** |

Table: Resultados obtenidos para la tarea AFC, considerando los mejores resultados de la Shared Task.

# Resultados para AFC

| Task | Method | Bagging | Text | Audio | Text + Audio |
|---|---|---|---|---|---|
| AFC | RoBERTa + Whisper† [Tahir et al., 2025] | ✗ | 0.486 | 0.159 | 0.461 |
| | RoBERTa + HuBERT† [Pittiglio, 2025] | | 0.444 | **0.356** | 0.440 |
| | (MM-)RoBERTa + WavLM† | | 0.393 | 0.064 | 0.382 |
| | Qwen2-Audio-7B (ICL) [Chu et al., 2024] | ✗ | – | – | 0.170 |
| | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | | – | – | 0.507 |
| | Llama3-3B+BYOL-A (SFT) | | 0.523 | – | – |
| | Our method | | – | – | 0.548 |
| | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | ✓ | – | – | 0.515 |
| | Llama3-3B+BYOL-A (SFT) | | **0.537** | – | – |
| | Our method | | – | – | **0.592** |

Table: Resultados obtenidos para la tarea AFC, considerando los mejores resultados de la Shared Task.

# Resultados para AFD

| Task | Method | Bagging | Text | Audio | Text + Audio |
|---|---|---|---|---|---|
| AFD | RoBERTa + wav2vec2.0† [Cantín and Chust, 2025] | ✗ | 0.220 | 0.169 | 0.193 |
| | (MM-)RoBERTa + WavLM† | | 0.277 | 0.000 | 0.285 |
| | Llama3-3B+BYOL-A (SFT) | ✗ | 0.283 | – | 0.285 |
| | Our method | | – | – | 0.307 |
| | Llama3-3B+BYOL-A (SFT) | ✓ | **0.297** | – | 0.301 |
| | Our method | | – | – | **0.317** |

Table: Resultados obtenidos para la tarea AFD, considerando los mejores resultados de la Shared Task.

# Resultados para AFD

| Task | Method | Bagging | Text | Audio | Text + Audio |
|---|---|---|---|---|---|
| AFD | RoBERTa + wav2vec2.0† [Cantín and Chust, 2025] | ✗ | 0.220 | 0.169 | 0.193 |
| | (MM-)RoBERTa + WavLM† | | 0.277 | 0.000 | 0.285 |
| | Llama3-3B+BYOL-A (SFT) | ✗ | 0.283 | – | 0.285 |
| | Our method | | – | – | 0.307 |
| | Llama3-3B+BYOL-A (SFT) | ✓ | **0.297** | – | 0.301 |
| | Our method | | – | – | **0.317** |

Table: Resultados obtenidos para la tarea AFD, considerando los mejores resultados de la Shared Task.

# Resultados para AFD

| Task | Method | Bagging | Text | Audio | Text + Audio |
|---|---|---|---|---|---|
| AFD | RoBERTa + wav2vec2.0† [Cantín and Chust, 2025] | ✗ | 0.220 | 0.169 | 0.193 |
| | (MM-)RoBERTa + WavLM† | | 0.277 | 0.000 | 0.285 |
| | Llama3-3B+BYOL-A (SFT) | ✗ | 0.283 | – | 0.285 |
| | Our method | | – | – | 0.307 |
| | Llama3-3B+BYOL-A (SFT) | ✓ | **0.297** | – | 0.301 |
| | Our method | | – | – | **0.317** |

Table: Resultados obtenidos para la tarea AFD, considerando los mejores resultados de la Shared Task.

# Resultados para SENT

| Task | Method | Bagging | Text | Text + Audio |
|---|---|---|---|---|
| SENT | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | ✗ | – | 0.769 |
| | Llama3-3B+BYOL-A (SFT) | | 0.759 | 0.763 |
| | Our method | | – | 0.774 |
| | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | ✓ | – | 0.779 |
| | Llama3-3B+BYOL-A (SFT) | | 0.764 | 0.771 |
| | Our method | | – | **0.792** |

Table: Resultados para la tarea SENT sobre CMU-MOSEI (Macro-F1).

# Resultados para SENT

| Task | Method | Bagging | Text | Text + Audio |
|---|---|---|---|---|
| SENT | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | | – | 0.769 |
| | Llama3-3B+BYOL-A (SFT) | ✗ | 0.759 | 0.763 |
| | Our method | | – | 0.774 |
| | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | | – | 0.779 |
| | Llama3-3B+BYOL-A (SFT) | ✓ | 0.764 | 0.771 |
| | Our method | | – | **0.792** |

Table: Resultados para la tarea SENT sobre CMU-MOSEI (Macro-F1).

# Resultados para EMO

| Task | Method | Bagging | Text | Text + Audio |
|---|---|---|---|---|
| EMO | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | ✗ | – | 0.731 |
| | Llama3-3B+BYOL-A (SFT) | | 0.719 | 0.728 |
| | Our method | | – | 0.736 |
| | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | ✓ | – | 0.734 |
| | Llama3-3B+BYOL-A (SFT) | | 0.726 | 0.732 |
| | Our method | | – | **0.742** |

Table: Resultados para la tarea EMO sobre CMU-MOSEI (Macro-F1).

# Resultados para EMO

| Task | Method | Bagging | Text | Text + Audio |
|---|---|---|---|---|
| EMO | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | ✗ | – | 0.731 |
| | Llama3-3B+BYOL-A (SFT) | | 0.719 | 0.728 |
| | Our method | | – | 0.736 |
| | Qwen2-Audio-7B (SFT) [Chu et al., 2024] | ✓ | – | 0.734 |
| | Llama3-3B+BYOL-A (SFT) | | 0.726 | 0.732 |
| | Our method | | – | **0.742** |

Table: Resultados para la tarea EMO sobre CMU-MOSEI (Macro-F1).

# Análisis sobre AFC

- El audio **siempre aporta** información complementaria (mejora incluso con selección *aleatoria*).

| Audio PT | Filtering Semantic | $\ell_1$ | F1 | |
|---|---|---|---|---|
| $\emptyset$ | $\emptyset$ | $\emptyset$ | 52.3 | Solo texto |
| ✗ | ✓ | ✗ | 48.7 | 7 capas + random |
| ✗ | ✓ | ✓ | 52.7 | 7 capas + $\ell_1$ |
| ✓ | ✗ | ✗ | 53.3 | 8 capas + random |
| ✓ | ✗ | ✓ | 51.8 | 8 capas + $\ell_1$ |
| ✓ | ✓ | ✗ | 53.9 | 7 capas + random |
| ✓ | ✓ | ✓ | **54.8** | 7 capas + $\ell_1$ |

Selección de tokens de audio en AFC (Macro-F1).

# Análisis sobre AFC

- El audio **siempre aporta** información complementaria (mejora incluso con selección *aleatoria*).
- Mejor usar **solo tokens acústicos**: el LLM ya cubre lo semántico del texto.

| Audio PT | Filtering Semantic | $\ell_1$ | F1 | |
|---|---|---|---|---|
| $\emptyset$ | $\emptyset$ | $\emptyset$ | 52.3 | Solo texto |
| ✗ | ✓ | ✗ | 48.7 | 7 capas + random |
| ✗ | ✓ | ✓ | 52.7 | 7 capas + $\ell_1$ |
| ✓ | ✗ | ✗ | 53.3 | 8 capas + random |
| ✓ | ✗ | ✓ | 51.8 | 8 capas + $\ell_1$ |
| ✓ | ✓ | ✗ | 53.9 | 7 capas + random |
| ✓ | ✓ | ✓ | **54.8** | 7 capas + $\ell_1$ |

Selección de tokens de audio en AFC (Macro-F1).

# Análisis sobre AFC

- El audio **siempre aporta** información complementaria (mejora incluso con selección *aleatoria*).
- Mejor usar **solo tokens acústicos**: el LLM ya cubre lo semántico del texto.
- Sin pre-entrenamiento, la **selección lasso ($\ell_1$)** es clave para estabilizar el rendimiento.

| Audio PT | Filtering Semantic | $\ell_1$ | F1 | |
|---|---|---|---|---|
| $\emptyset$ | $\emptyset$ | $\emptyset$ | 52.3 | Solo texto |
| ✗ | ✓ | ✗ | 48.7 | 7 capas + random |
| ✗ | ✓ | ✓ | 52.7 | 7 capas + $\ell_1$ |
| ✓ | ✗ | ✗ | 53.3 | 8 capas + random |
| ✓ | ✗ | ✓ | 51.8 | 8 capas + $\ell_1$ |
| ✓ | ✓ | ✗ | 53.9 | 7 capas + random |
| ✓ | ✓ | ✓ | **54.8** | 7 capas + $\ell_1$ |

Selección de tokens de audio en AFC (Macro-F1).

# Tamaño del vocabulario de audio

- Tokens de audio **seleccionados**: **73** (config. 7 capas) y **80** (config. 8 capas).
- En ambos casos, **menos del 1%** del vocabulario inicial.

$$\frac{|\mathcal{V}'_a|}{|\mathcal{V}_a|} \approx 10^{-2}$$

- Largo promedio de secuencia $\rightarrow$ Texto: **31.52** | Audio: **35.52** (mismo orden de magnitud).

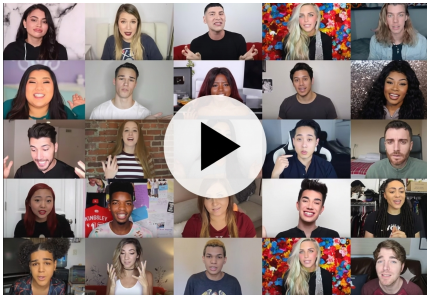
**Idea clave**

Seleccionar los tokens acústicos relevantes mantiene la secuencia de audio **comparable a la del texto**, usando $< 1\%$ del vocabulario.

# Análisis cualitativo SENT/EMO

## Hallazgo (CMU-MOSEI)

El modelo multimodal usa las **señales acústicas** cuando el texto no basta para desambiguar la clase: **texto neutro** dicho con **voz alegre o triste**.



Clip de CMU-MOSEI — youtu.be/QvjJg_GjcvE

## Ejemplo

*"Do you need to wait to be admitted to apply for financial aid?"*

- Solo texto → negativo / sin emoción
- Texto + audio → positivo / happy

El texto es neutro: la voz aporta el afecto.

# Conclusiones, limitaciones y trabajo futuro

# Conclusiones

## Contribuciones

- Método simple para **integrar tokens de audio** en un LLM textual, manejando su alta frecuencia mediante **selección de tokens relevantes** (regresión logística lasso).

# Conclusiones

## Contribuciones

- Método simple para **integrar tokens de audio** en un LLM textual, manejando su alta frecuencia mediante **selección de tokens relevantes** (regresión logística lasso).
- Evaluación del rendimiento:
  - de modelos "básicos" de Machine Learning usando representaciones textuales y auditivas.
  - de LLMs (texto y texto+audio) en las tareas AFC, AFD, SENT y EMO.

# Conclusiones

## Contribuciones

- Método simple para **integrar tokens de audio** en un LLM textual, manejando su alta frecuencia mediante **selección de tokens relevantes** (regresión logística lasso).
- Evaluación del rendimiento:
  - de modelos "básicos" de Machine Learning usando representaciones textuales y auditivas.
  - de LLMs (texto y texto+audio) en las tareas AFC, AFD, SENT y EMO.
- Análisis en profundidad de las configuraciones de selección de tokens de audio.

# Conclusiones

## Conclusión

- La integración de información auditiva mejora el rendimiento de los LLMs en tareas donde el audio se creía **no útil**, alcanzando resultados **estado del arte** frente a baselines fuertes, incluido un **modelo multimodal más grande**.

## Producción

- Artículo "A Simple Method to Enhance Pre-trained Language Models with Speech Tokens for Classification", propuesto a EACL 2026 (pendiente de revisión). Preprint: arXiv:2512.07571.
- Código y modelos disponibles en GitHub.

# Limitaciones y Trabajo futuro

- **Limitaciones**:
  - Los experimentos se restringen a un escenario de **clasificación** (no generación).
  - Los embeddings de audio se ajustan **junto a la tarea**; no se evaluó su generalización sin ese fine-tuning.
  - Solo se entrenan las capas de **embedding** (vía causal language modeling), no todos los parámetros.

# Limitaciones y Trabajo futuro

- **Limitaciones**:
  - Los experimentos se restringen a un escenario de **clasificación** (no generación).
  - Los embeddings de audio se ajustan **junto a la tarea**; no se evaluó su generalización sin ese fine-tuning.
  - Solo se entrenan las capas de **embedding** (vía causal language modeling), no todos los parámetros.
- **Trabajo futuro**:
  - Extender el método a tareas de **generación de lenguaje**.
  - Probar una **menor regularización lasso** para obtener más tokens y hallar el tamaño óptimo.
  - Estudio **correlacional** entre los tokens aprendidos y features acústicas expertas (pitch, intensidad, forma espectral, flujo glótico) para mejorar la interpretabilidad.
  - Análisis del **flujo de atención** (cómo el modelo integra la información acústica a través de las capas).

# A Simple Method to Enhance Pre-trained Language Models with Speech Tokens for Classification

Nicolas Calbucura **José Guillen** Valentin Barriere

*En proceso de revisión en EMNLP 2026*
*Preprint: arXiv:2512.07571*

3 de Junio, 2026

# Referencias I


Cantín, E. and Chust, A. (2025).
Argumentative Fallacy Detection in Political Debates.
In Chistova, E., Cimiano, P., Haddadan, S., Lapesa, G., and Ruiz-Dolz, R., editors, *Proceedings of the 12th Argument mining Workshop*, pages 369–373, Vienna, Austria. Association for Computational Linguistics.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., and Zhou, J. (2024).
Qwen2-Audio Technical Report.
pages 1–16.

# Referencias II


Fröhlich, M., Sievers, C., Townsend, S. W., Gruber, T., and van Schaik, C. P. (2019).
Multimodal communication and language origins: integrating gestures and vocalizations.
*Biological Reviews*, 94(5):1809–1829.

Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., and Kashino, K. (2023).
BYOL for Audio: Exploring Pre-Trained General-Purpose Audio Representations.
*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:137–151.

Pittiglio, A. (2025).
Leveraging Context for Multimodal Fallacy Classification in Political Debates.
In Chistova, E., Cimiano, P., Haddadan, S., Lapesa, G., and Ruiz-Dolz, R., editors, *Proceedings of the 12th Argument mining Workshop*, pages 388–397, Vienna, Austria. Association for Computational Linguistics.

# Referencias III


Tahir, A., Ibrar, I., Ameer, H., Fatima, M., and Latif, S. (2025).
Prompt-Guided Augmentation and Multi-modal Fusion for Argumentative Fallacy Classification in Political Debates.
In Chistova, E., Cimiano, P., Haddadan, S., Lapesa, G., and Ruiz-Dolz, R., editors, *Proceedings of the 12th Argument mining Workshop*, pages 381–387, Vienna, Austria. Association for Computational Linguistics.

Thimonier, H., Perzo, A., and Seguier, R. (2025).
EmoSLLM: Parameter-Efficient Adaptation of LLMs for Speech Emotion Recognition.
pages 1–18.