## Capítulo 1 – Introducción

### 1.1 Justificación y objetivo del trabajo

Hablar en público constituye una competencia clave en contextos académicos, profesionales y sociales. La efectividad comunicativa en estos escenarios no depende únicamente del contenido verbal, sino también de elementos paralingüísticos (como las pausas, el tono y el ritmo) y no verbales (como los gestos, la postura o la expresividad facial), que inciden directamente en la claridad, la credibilidad y el impacto del mensaje.

En la actualidad, con la creciente expansión de formatos digitales como videollamadas, clases virtuales y presentaciones en línea, el video se ha consolidado como un medio idóneo para estudiar la oratoria y los estilos comunicativos. Sin embargo, muchas personas siguen sin conocer qué patrones específicos mejoran una presentación ni cómo pueden observarse de manera objetiva y sistemática. A pesar de su relevancia, la comunicación efectiva sigue siendo una habilidad poco analizada con herramientas accesibles y basadas en datos.

Este trabajo responde a esa necesidad mediante el desarrollo de un enfoque de análisis multimodal, que examina discursos grabados considerando tres dimensiones principales:

* Dimensión textual: estructura argumentativa y relevancia del contenido.
* Dimensión acústica: ritmo de habla, tono, volumen y pausas.
* Dimensión visual: postura corporal, gesticulación y expresividad facial.

Para ello, los videos son segmentados en fragmentos breves, lo que permite un análisis granular y progresivo. A partir de cada segmento se extraen características cuantificables que alimentan modelos de aprendizaje automático capaces de detectar patrones asociados a momentos clave del discurso y a la calidad percibida del mensaje.

El objetivo general de este trabajo es doble:

1. Comprender qué combinaciones de elementos multimodales favorecen una comunicación más efectiva.
2. Contribuir al desarrollo de herramientas que permitan a cualquier persona evaluar y mejorar sus presentaciones de forma autónoma, promoviendo el autoaprendizaje estructurado y basado en datos reales.

Todo el sistema ha sido implementado como un conjunto de notebooks en Google Colab, lo que permite su ejecución directa desde la nube sin necesidad de instalaciones locales.

El cuaderno principal (6\_produccion.ipynb) permite evaluar vídeos individuales aplicando el modelo entrenado y generando retroalimentación automática.

📁 Acceso completo al sistema: https://drive.google.com/drive/folders/15LFR3rK3jT\_EtLpjtoeKOWTUoHEcF-1d?usp=drive\_link  
📒 Notebook principal: [enlace a 6\_produccion.ipynb]

### 1.2 Estructura del trabajo

El presente trabajo se estructura en seis capítulos:

* **Capítulo 1. Introducción**, donde se presenta el contexto, la motivación y los objetivos generales del estudio.
* **Capítulo 2. Marco teórico**, que revisa los conceptos clave sobre análisis del discurso, comunicación multimodal, detección automática de emociones y aprendizaje automático aplicado a la comunicación oral.
* **Capítulo 3. Construcción y limpieza del dataset multimodal**, donde se documenta el proceso completo de extracción, segmentación, análisis multimodal, depuración de variables y creación del dataset unificado de entrenamiento.
* **Capítulo 4. Modelado predictivo y generación de feedback**, en el que se entrena un modelo explicable (Random Forest) para estimar la calidad comunicativa de un discurso, se evalúa su rendimiento y se implementa un sistema de retroalimentación basado en características multimodales.
* **Capítulo 5. Prueba de producción,** que valida el sistema en un entorno ejecutable (Google Colab), con funciones automatizadas para procesar nuevos vídeos y generar resultados de forma reproducible.
* **Capítulo 6. Conclusiones y líneas futuras**, donde se presentan las principales conclusiones del trabajo, se reflexiona sobre los resultados obtenidos y se proponen posibles mejoras y aplicaciones futuras en contextos educativos, profesionales o personales.

## Capítulo 2 – Marco Teórico

### 2.1 Análisis del discurso oral

El discurso oral es una forma de comunicación compleja que combina el contenido lingüístico con elementos paralingüísticos y no verbales. A diferencia del lenguaje escrito, el habla se produce en tiempo real y está profundamente influida por el contexto, lo que exige una planificación estructural que favorezca la claridad y el impacto del mensaje. La mayoría de los discursos efectivos comparten una organización en tres partes: introducción, desarrollo y cierre. Además del contenido, aspectos como el ritmo, las pausas y el énfasis desempeñan un papel clave en la atención y comprensión del oyente. Estos elementos, tradicionalmente estudiados desde la lingüística y la retórica, adquieren hoy una nueva relevancia en el análisis automatizado del discurso.

### 2.2 Comunicación multimodal

La comunicación en contextos orales no se limita al lenguaje verbal, sino que integra componentes auditivos y visuales que modulan o refuerzan el mensaje. Este enfoque multimodal contempla la interacción entre lo que se dice (texto), cómo se dice (voz) y cómo se acompaña (gestos, posturas, expresiones faciales). Elementos como el tono de voz, las pausas o la expresividad corporal influyen en la interpretación emocional y en la percepción de credibilidad. Por ello, un análisis completo del discurso debe considerar todos estos canales como parte de un mismo sistema expresivo.

### 2.3 Detección de emociones en audio y video

Las emociones forman parte integral de la comunicación y pueden manifestarse a través de la voz, el rostro y el lenguaje corporal. En el ámbito computacional, la detección emocional se basa en modelos categoriales (como el de Ekman, centrado en emociones básicas) o dimensionales (como el de valencia y activación). El análisis se realiza a partir de señales acústicas (variaciones de tono, ritmo o energía), visuales (microexpresiones, sonrisas) y textuales (contenido semántico). La combinación de estos canales —análisis multimodal— permite una detección más precisa y contextualizada de los estados emocionales presentes en un discurso.

### 2.4 Procesamiento de audio, texto y video

Cada canal del discurso requiere un tratamiento técnico específico. En audio, se emplean características como MFCC, pitch, energía o tasa de pausas. El texto, generado mediante transcripción automática, permite analizar la estructura retórica y el contenido semántico. En video, se extraen posturas, gestos y expresiones faciales mediante técnicas de visión por computador. La integración de estos datos ofrece una representación rica del acto comunicativo, clave para su posterior análisis mediante modelos de aprendizaje automático.

### 2.5 Machine Learning en análisis multimodal

El aprendizaje automático permite descubrir patrones en grandes volúmenes de datos multimodales. Se utilizan desde modelos clásicos como Random Forest hasta arquitecturas neuronales más complejas como CNN, LSTM o transformers. La fusión de datos de distintos canales mejora la capacidad predictiva, aunque plantea desafíos como la sincronización temporal o la variabilidad individual. En este trabajo, los modelos se utilizan para identificar momentos clave del discurso y evaluar su efectividad comunicativa.

### 2.6 Trabajos relacionados

Diversos estudios han abordado el análisis del discurso desde perspectivas textuales, acústicas o visuales, aunque muchos lo hacen de forma aislada o en contextos controlados. Propuestas como Presentation Trainer o competiciones como ComParE han impulsado el desarrollo de modelos aplicados a la voz o al gesto, pero pocas soluciones son ligeras y adaptables a escenarios reales. Este trabajo propone una aproximación integrada, accesible y centrada en presentaciones reales, combinando análisis discursivo, emocional y gestual para facilitar el autoaprendizaje y la mejora comunicativa.

## Capítulo 3 – Construcción del dataset multimodal.

### 3.1 Selección de vídeos para la construcción del dataset

#### 3.1.1 Fuente y criterios de selección

Uno de los principales retos para este estudio fue la disponibilidad de discursos orales reales en formato audiovisual que permitieran un análisis multimodal. Como fuente principal se utilizó el conjunto de datos disponible en la plataforma Kaggle bajo el título **“TED Talks”**, que incluye **5.440 vídeos** de presentaciones acompañadas de metadatos relevantes: título, autor, fecha, visualizaciones, "likes" y enlace al vídeo original..

El conjunto de datos está disponible en:

<https://www.kaggle.com/datasets/ashishjangra27/ted-talks>

Este recurso se distribuye bajo la licencia **Creative Commons BY-NC-SA 4.0**, que permite su uso en proyectos académicos y no comerciales siempre que se cite adecuadamente.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

#### 3.1.2 Definición de clases y agrupamiento por lotes

Con el objetivo de evaluar la calidad comunicativa de los oradores, se utilizó el número de **“likes”** como una **aproximación indirecta y cuantitativa** a la calidad percibida por la audiencia.

Para facilitar el análisis comparativo y reducir la carga computacional del procesamiento multimodal, se seleccionaron únicamente los extremos del conjunto:

El **25% de los vídeos con mayor número de likes** (n=1370), fueron etiquetados como **buenos comunicadores (clase 1)**.

El **25% con menor número de likes** (n=1400), fueron etiquetados como **malos comunicadores (clase 0)**.

Esta estrategia responde a varios motivos:

* Mayor **separabilidad de clases**: se trabaja con ejemplos claros.
* Menor **ambigüedad**: se evita el ruido de casos intermedios.
* Mayor **eficiencia computacional**: se reduce el volumen total de vídeos al 50%.
* Tamaño **suficiente y equilibrado** para el entrenamiento de modelos.

Dado el alto coste computacional del procesamiento, los **2.770 vídeos seleccionados** se organizaron en **100 grupos balanceados**, asegurando que cada lote contuviera vídeos de ambas clases. Esto facilitó el procesamiento por lotes en entornos como Google Colab, donde los recursos pueden ser limitados.

#### 3.1.3 Uso de la probabilidad como medida continua

Además de la clasificación binaria, **se pretende utilizar** la **probabilidad de pertenencia a la clase 1** —calculada por los modelos de aprendizaje automático— como una **medida continua de calidad comunicativa**. Esta estrategia permite:

* Ordenar los discursos según su **calidad percibida**.
* Capturar la **incertidumbre** del modelo más allá del 0/1.
* Calibrar mejor la distribución real de niveles de calidad.
* Evaluar discursos intermedios que no fueron utilizados en el entrenamiento.

Los cálculos técnicos de esta estrategia se detallan en el **Anexo 1:**

**1\_seleccion\_videos\_conjunto\_de\_datos**.

Creamos así el archivo “videos\_analisis\_csv” que con tiene 2770 datos de videos con el campo "categoria" 0 y 1 que posteriormente se utilizarán para el entrenamiento del modelo y renombrando el campo como "tipo\_comunicador"

Distribución de los datos en el archivo de estudio:

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

### 3.2 Procesamiento de los vídeos y construcción del dataset

Las operaciones descritas en este apartado se implementaron en Python y pueden consultarse en el **Anexo 3:**

**3\_extraccion\_datos\_videos\_ted.ipynb**

#### 3.2.1 Pipeline general

La construcción del dataset se realizó mediante un **proceso modular automatizado**, diseñado para extraer características relevantes a partir del contenido audiovisual de los vídeos seleccionados. El pipeline general siguió las siguientes etapas:

1. **Agrupación de vídeos** en 100 lotes con mezcla equilibrada de tipos de comunicador.
2. **Descarga automatizada** de cada vídeo (audio + video) usando yt-dlp.
3. **Segmentación dinámica** del discurso en función de pausas y cambios acústicos.
4. **Transcripción del audio** mediante faster-whisper, con alineación temporal precisa.
5. **Extracción de características acústicas** con librosa y torchaudio.
6. **Análisis visual** a partir de fotogramas centrales de cada segmento, usando cv2, YOLO y modelos ligeros de detección de rostro y posturas.
7. **Integración multimodal** Se serializan las variables de audio, texto y vídeo alineadas temporalmente en JSON por vídeo y en JSON por lote (grupo). Una vez procesados todos los lotes, se integran en un repositorio unificado y en formatos tabulares (CSV) para el modelado y el análisis estadístico.

#### 3.2.2 Segmentación dinámica del discurso

Cada vídeo fue dividido en **segmentos dinámicos** de duración variable (media aproximada de 2.5 segundos), en función de **pausas naturales** y **cambios acústicos** que indicaran posibles variaciones en el énfasis o la intención comunicativa del hablante.

Este tipo de segmentación permite centrar el análisis en los momentos más expresivos o relevantes del discurso, facilitando la detección de **patrones emocionales, estructurales o retóricos**. A diferencia de enfoques basados en bloques predefinidos (como introducción o conclusión), esta estrategia se adapta al ritmo real del orador.

#### 3.2.3 Extracción y organización de características por segmento y por vídeo

Durante el procesamiento de cada vídeo, se extrajeron y organizaron **características multimodales** a nivel de **vídeo completo** y de **segmento individual**, con el objetivo de capturar tanto el contexto general del discurso como los matices expresivos específicos de cada momento.

**🔹 A nivel de vídeo**

* **video\_id**: identificador único del vídeo.
* **link**: URL original del vídeo en TED.
* **tipo\_comunicador**: clase binaria asignada al orador: 0 (baja calidad) o 1 (alta calidad).
* **grupo**: número del grupo de procesamiento al que pertenece el vídeo (0–99).
* **duracion\_video**: duración total del vídeo en segundos.
* **texto\_completo**: transcripción completa del discurso, generada automáticamente.
* **idioma**: idioma detectado en la transcripción.

**🔹 A nivel de segmento**

Cada vídeo fue segmentado dinámicamente y, para cada uno de estos fragmentos, se recopilaron las siguientes variables:

**📄 Información general**

* **seg\_id**: identificador único del segmento dentro del vídeo.
* **texto**: transcripción del contenido del segmento.
* **inicio / fin**: tiempos de inicio y finalización del segmento (en segundos).
* **duracion**: duración total del segmento.
* **pausa\_anterior**: duración de la pausa anterior al inicio del segmento.

**🔊 Características acústicas**

* **rms\_mean**: energía promedio del audio (Root Mean Square).
* **zcr\_mean**: tasa de cruce por cero, relacionada con la sonoridad.
* **tipo\_audio**: clasificación del sonido predominante (voz, silencio, ruido).
* **pmm**: palabras por minuto estimadas en el segmento.
* **emocion\_audio**: emoción inferida a partir de un modelo entrenado con el conjunto de datos RAVDESS. Este modelo será descrito en detalle en consideraciones técnicas

**🎥 Análisis visual (imagen central del segmento)**

* **t\_central**: instante temporal central del segmento.
* **frame\_id**: identificador del fotograma utilizado para el análisis visual.
* **cara\_detectada**: detección automática de rostro en el fotograma (sí/no).
* **yaw / pitch / roll**: orientación de la cabeza en los tres ejes espaciales.
* **boca\_abierta**: grado de apertura de la boca.
* **sonrisa / sonrisa\_detectada**: intensidad y presencia de sonrisa.
* **ceño\_fruncido / ceño\_detectado**: intensidad y detección binaria del ceño fruncido.
* **ojos\_abiertos**: nivel de apertura ocular.
* **asimetria\_labios**: diferencia de posición entre los extremos de la boca.
* **tension\_facial**: estimación de tensión muscular en el rostro.
* **estado\_emocional**: emoción visual inferida (alegría, tristeza, neutro, etc.).

**🧍‍♂️ Postura corporal y manos**

* **apertura\_brazos**: grado de extensión de los brazos, asociado a la expresividad.
* **inclinacion\_torso**: inclinación del torso respecto al eje vertical.
* **manos\_visibles**: Detección de las manos en el fotograma (sí/no)
* **mano\_izq\_x / mano\_izq\_y**: coordenadas normalizadas de la mano izquierda.
* **mano\_der\_x / mano\_der\_y**: coordenadas normalizadas de la mano derecha.

Todas estas variables se almacenan en estructuras JSON organizadas por vídeo y por grupo, para su posterior integración en una base de datos multimodal completa y reutilizable.

#### 3.2.5 Consideraciones técnicas

Todo el procesamiento se llevó a cabo en entornos controlados, principalmente utilizando **Google Colab** con detección automática de **aceleración por GPU (CUDA)** cuando estuvo disponible. Dado el elevado volumen de datos y la naturaleza multimodal del análisis, se priorizó el uso de herramientas eficientes y estructuras de procesamiento por lotes.

Para asegurar una ejecución fluida y replicable, se diseñó un sistema modular que permite:

* Ejecutar el procesamiento por grupos de vídeos (100 grupos balanceados).
* Guardar resultados intermedios en formato **JSON estructurado**, por vídeo y por grupo.

El pipeline utilizó las siguientes bibliotecas principales:

* yt-dlp: descarga automatizada de vídeos desde TED.
* faster-whisper: transcripción automática de audio con alineación temporal precisa.
* librosa y torchaudio: análisis de características acústicas.
* opencv (cv2), YOLO, PoseEstimationWithMobileNet: análisis visual por fotograma.
* pandas, numpy, joblib, torch, moviepy: procesamiento y manipulación general de datos.

##### Modelo de detección de emociones acústicas

Las operaciones descritas para la creación de este modelo se implementaron en Python y pueden consultarse en el **Anexo 2:**

**2\_Modelo\_emociones\_audio\_Ravdess.ipynb**.

Para estimar la **emoción dominante** en cada segmento de audio, se desarrolló un modelo supervisado entrenado con el conjunto de datos **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)**, un corpus de referencia en el ámbito del análisis emocional. Este dataset incluye grabaciones de voz actuadas con etiquetas emocionales discretas, tales como *alegría, tristeza, ira, miedo, calma, sorpresa*, entre otras.

- Dataset original disponible en Zenodo:

🔗 https://zenodo.org/record/1188976

📄 Licencia: [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)

- La versión utilizada fue descargada desde Kaggle (espejo no oficial):

🔗 https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

Este dataset y los modelos derivados están destinados únicamente a fines académicos.

Cualquier reutilización debe respetar los términos de la licencia original, incluyendo la obligación de compartir con la misma licencia (SA).

El proceso de construcción del modelo siguió las siguientes etapas:

* **Etiquetado automático:** los archivos .wav de RAVDESS fueron organizados y etiquetados a partir de los códigos incluidos en sus nombres.
* **Extracción de características acústicas**: para cada grabación se extrajeron y promediaron:
  + **ZCR (Zero Crossing Rate)**: indicador de sonoridad.
  + **RMS (Root Mean Square)**: medida de energía.
  + **MFCC (13 coeficientes)**: representación espectral compacta.
  + **Mel Spectrogram** (resumen): energía distribuida en bandas perceptuales.

Gráfico

El contenido generado por IA puede ser incorrecto.Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Antes del entrenamiento, las emociones etiquetadas en cada archivo de audio fueron **codificadas numéricamente** mediante un LabelEncoder, transformando las categorías textuales (por ejemplo, *"alegría"*, *"tristeza"*) en valores enteros. Esta codificación es necesaria para que los modelos de clasificación puedan operar sobre variables categóricas.

Las características acústicas extraídas fueron **normalizadas** utilizando un StandardScaler, con el objetivo de escalar todas las variables a una distribución común (media cero y desviación estándar uno). Esta etapa es fundamental para garantizar una convergencia eficiente durante el entrenamiento y evitar que las variables de mayor magnitud dominen el proceso de aprendizaje.

Tanto el **escalador** (scaler.pkl) como el **codificador de etiquetas** (label\_encoder.pkl) fueron guardados para ser reutilizados en la predicción sobre segmentos reales del corpus TED, garantizando la coherencia entre el entrenamiento y la inferencia.

El modelo se implementó mediante una **red neuronal simple con Keras (TensorFlow)**. Su arquitectura incluyó:

* Capas densas con activación ReLU y regularización L2.
* Capas ocultas con Dropout para controlar el sobreajuste.
* Capa de salida softmax, con una neurona por clase emocional.

Durante el entrenamiento se aplicaron técnicas como **early stopping** y **reduce-on-plateau** para optimizar la convergencia y evitar el sobreajuste.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

Una vez entrenado, se guardó el modelo (deep\_model.pth), que posteriormente se aplicó a los segmentos del corpus TED, utilizando el mismo pipeline de extracción de características. El resultado se incorporó como la variable **emocion\_audio**, que permite enriquecer el análisis comunicativo con un estimador del tono emocional predominante en cada segmento. Este valor **no se utiliza como variable dependiente**, sino como **apoyo contextual** para identificar momentos expresivos o afectivamente cargados dentro del discurso oral.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

### 3.3 Limpieza y estructuración del dataset

Tras la extracción multimodal por vídeo y por segmento, se aplicó un proceso de limpieza, integración y estandarización con el fin de preparar los datos para el análisis y el modelado.

Las operaciones descritas en este apartado se implementaron en Python y pueden consultarse en el **Anexo 4:**

**4\_limpieza\_estructuracion\_dataset\_final.ipynb**.

#### 3.3.1 Unificación de archivos JSON

Cada grupo de vídeos procesado generó un archivo independiente en formato **.json** con la información estructurada de cada vídeo y sus segmentos. Estos ficheros se integraron en un único archivo consolidado, **features\_videos\_ted.json**, almacenado en la carpeta de trabajo principal.  
El fichero unificado contiene, por cada video\_id, los **metadatos generales** (idioma, duración, clase/categoría, texto completo) y la **lista de segmentos** analizados.

#### 3.3.2 Construcción de datasets tabulares

A partir del JSON unificado se generaron dos estructuras tabulares:

* **a) Dataset a nivel de vídeo** (df\_videos\_analizados.csv): una fila por vídeo. Para simplificar su estructura, se eliminó la columna segmentos, tratada por separado. Incluye:

video\_id, link, tipo\_comunicador, grupo, duracion\_video, texto\_completo, idioma.

Obtenemos un total de 2.669 videos, de los cuales 1.349 son “tipo\_comunicador” 0 (mal comunicador) y 1.320 son “tipo\_comunicador” 1 (buen comunicador). Todos sin nulos en estos valores. La duración de estos videos es desde los 60 segundos a los 340 minutos con una media de 12 minutos de duración. El idioma principal es el inglés con 2540 videos.

La diferencia entre los 2.669 videos que hemos analizado y los 2.770 videos del listado inicial que pretendíamos analizar, se debe a errores en la descarga y análisis de estos.

* **b) Dataset a nivel de segmento**: una fila por cada fragmento del discurso, reorganizando variables **textuales**, **acústicas** y **visuales** desde el JSON. Incluye:
  + **Identificadores:** video\_id, seg\_id.
  + **Texto:** texto, pmm.
  + **Audio:** rms\_mean, zcr\_mean, tipo (Habla/Pausa), emocion\_audio.
  + **Visual:** (expresiones faciales, postura, manos, inclinaciones).
  + **Tiempo:** inicio, fin, duracion.

Además, se **eliminó el tipo de segmento “Pausa”**, se **reordenaron** los segmentos, se **reasignó** el número seg\_id y se **recalculó** la pausa\_anterior para reflejar la nueva secuencia.

Obtenemos así dataset con un total de 860.632 segmentos correspondientes a los 2.669 videos.

#### 3.3.3 Normalización semántica y depuración de variables

A partir de este punto, el proceso de limpieza y depuración se centra exclusivamente en el **dataset a nivel de segmento**, ya que es en esta unidad donde se concentran las variables multimodales objeto de análisis (textuales, acústicas y visuales). El dataset a nivel de vídeo, por su estructura más simple, no requiere transformación adicional.

Se comenzó por depurar el dataset de segmentos, identificando variables problemáticas, redundantes o mal definidas, y eliminándolas del análisis para asegurar la consistencia semántica del conjunto de datos.

Las principales decisiones de depuración fueron las siguientes:

* **Manos (coordenadas):** las variables mano\_izq\_x/y y mano\_der\_x/y estaban referidas a la posición absoluta en la imagen, y no a un marco relativo al rostro o al cuerpo, por lo que sus valores no eran comparables entre vídeos. Se eliminaron. Se mantuvo la variable **apertura\_brazos** como indicador representativo de gesticulación.
* **Estado emocional visual:** se eliminó la variable compuesta estado\_emocional, generada automáticamente a partir de otras variables como sonrisa, ceño\_fruncido, tension\_facial, etc. Su eliminación evita duplicidades y permite centrarse en variables continuas directamente observables.
* **Redundancias booleanas:** las variables ceño\_detectado y sonrisa\_detectada, de tipo booleano, se eliminaron por ser redundantes respecto a sus contrapartes continuas: ceño\_fruncido y sonrisa, respectivamente. Esta decisión también contribuye a reducir la dimensionalidad del dataset sin pérdida de información relevante.

Estas operaciones permitieron alinear semánticamente el dataset, priorizando variables primitivas y medibles, con mayor utilidad para el modelado y la interpretación de resultados.

#### 3.3.4 Estudio de outliers y winsorización

Se exploró la presencia de **valores atípicos** mediante funciones de visualización específicas y estadísticas descriptivas.

Algunos ejemplos:

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.Imagen que contiene Histograma

El contenido generado por IA puede ser incorrecto.

Reviso también las categóricas

Interfaz de usuario gráfica, Aplicación, Teams

El contenido generado por IA puede ser incorrecto.

Dado el volumen y la heterogeneidad del corpus, se optó por **winsorizar** las variables continuas (acústicas y visuales), con el fin de **reducir la influencia de extremos** sin eliminar observaciones. La winsorización permite **robustecer las estimaciones** (medias/varianzas), **preservar el rango efectivo** de las variables y **evitar sesgos** por eliminación masiva de datos. Los **límites aplicados** se almacenaron en **winsor\_limits.csv** para su trazabilidad y replicación.

#### 3.3.5 Análisis de valores ausentes

El dataset de segmentos alcanzó un total de **860.632 filas**. Se observó que el **50,90%** de ellas contenían al menos un valor nulo.

Imagen que contiene guitarra

El contenido generado por IA puede ser incorrecto.

El análisis estratificado por “cara\_detectada” mostró que:

* **Con cara detectada:** **3,98%** de las filas presentan nulos, principalmente en apertura\_brazos e inclinacion\_torso.
* **Sin cara detectada:** **100%** e las filas tienen nulos en variables visuales derivadas del rostro: yaw, pitch, roll, boca\_abierta, sonrisa, ceño\_fruncido, ojos\_abiertos, tension\_facial. Lo esperable al no detectarse el rostro.

#### 3.3.6 Filtro por cobertura facial a nivel de vídeo

Para mejorar la calidad del análisis visual sin perder una proporción excesiva de datos, se calculó, para cada vídeo, el **porcentaje de segmentos en los que se detectó rostro** (cara\_detectada = True). Esta proporción se agrupó en **10 intervalos** (*bins*) regulares del 0% al 100%, y se analizaron los recuentos acumulados para determinar en qué punto se lograba el mejor equilibrio entre calidad visual y tamaño del dataset.

Gráfico, Gráfico de barras, Histograma

El contenido generado por IA puede ser incorrecto.

Para decidir donde hacemos el corte y con cuantos videos nos quedamos hacemos revisamos acumulativamente.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Finalmente, se optó por conservar únicamente los vídeos en los que al menos el 30% de sus segmentos tienen detección facial positiva. Este umbral garantiza una cobertura visual mínima suficiente para el análisis de variables como gesticulación, expresividad o postura, sin reducir drásticamente el tamaño total del corpus.

Esta decisión representa un compromiso entre:

* Evitar imputaciones excesivas en vídeos sin rostro visible, y
* Mantener un conjunto suficientemente amplio y representativo.

Como resultado, el corpus se redujo a **1.936 vídeos**, todos con una proporción adecuada de segmentos con cara visible.

Como efecto colateral, este filtro también eliminó los casos en los que la variable frame\_id era nula. Se mantuvo la variable cara\_detectada en el dataset final como **marcador auxiliar de calidad visual**, útil para auditorías o análisis estratificados en etapas posteriores.

#### 3.3.7 Imputación de valores faltantes

Dado que **no era viable** descartar todas las filas con nulos, se definió una **estrategia de imputación robusta**:

* **Imputación por mediana *intra-vídeo***: para **cada columna numérica**, los valores faltantes se **reemplazaron por la mediana calculada dentro de cada video\_id**.

Este enfoque mantiene la **coherencia interna** de cada discurso (voz, sala, encuadre, estilo gestual), evitando introducir sesgos entre vídeos con condiciones distintas.

Tras aplicar esta técnica, persistían valores nulos en algunos segmentos pertenecientes a **tres vídeos concretos**. Dado su escaso volumen y la imposibilidad de completarlos adecuadamente, se optó por **eliminar esos vídeos en su totalidad**.

El resultado final fue un conjunto de **1.933 vídeos con datos completos a nivel de segmento**.

Este dataset definitivo, ya procesado mediante winsorización e imputación, fue almacenado en el archivo **df\_seg\_win\_med.csv**, que será utilizado en las siguientes etapas de análisis y modelado.

#### 3.3.8 Unificación de la información entre datasets.

Tras el proceso de limpieza y filtrado, quedaron definidos dos datasets principales:

* Un **dataset a nivel de vídeo**, con información general por video\_id.
* Un **dataset a nivel de segmento**, con información detallada por fragmento del discurso.

Para asegurar la consistencia entre ambos, se procedió a **eliminar del dataset de vídeos aquellos video\_id que ya no estaban presentes en el dataset de segmentos**, como consecuencia del filtrado y la imputación aplicados previamente.

El punto de partida para esta integración fue el conjunto de **1.933 vídeos** con datos completos a nivel de segmento, resultado final de la limpieza detallada en los apartados anteriores.

A continuación, se revisó la variable **idioma**, presente en el dataset de vídeos. De los 1.933 vídeos disponibles, se decidió conservar únicamente aquellos en **inglés (1.819 vídeos)** y **castellano (23 vídeos)**, descartando otros idiomas con escasa representación. Esta decisión se tomó para **facilitar los análisis lingüísticos y de contenido textual**.

El conjunto final quedó compuesto por **1.852 vídeos** con datos consistentes tanto a nivel global como segmentado, distribuidos de la siguiente forma según su clase:

* **955 vídeos** de tipo comunicador **0** (baja calidad comunicativa)
* **897 vídeos** de tipo comunicador **1** (alta calidad comunicativa)

### 3.4 Cálculo de características adicionales

En este apartado se generaron variables adicionales destinadas a enriquecer la representación de cada discurso a nivel de vídeo, complementando las características multimodales extraídas por segmento.

Las operaciones descritas en este apartado, incluyendo el análisis textual y la agregación de variables por vídeo, se implementaron en Python y pueden consultarse en el **Anexo 4 – limpieza\_estructuracion\_dataset\_final.ipynb**.

#### **3.4.1 Características textuales a nivel de vídeo**

Se aplicó un análisis semántico y retórico sobre la transcripción completa de cada discurso (texto\_completo), utilizando funciones lingüísticas, patrones heurísticos y modelos de embeddings multilingües (MiniLM). A partir de este proceso se extrajeron cinco variables clave:

* Número de oraciones detectadas.
* Uso de palabras de impacto emocional o persuasivo.
* Presencia de anécdotas personales.
* Inclusión de ejemplos y analogías.
* Frecuencia de expresiones en primera persona.

Estas métricas permiten cuantificar dimensiones relevantes del discurso como claridad, narrativa e intención persuasiva, y fueron incorporadas al dataset unificado para modelado.

Adicionalmente, se evaluó la correlación entre estas variables y la clase objetivo (tipo\_comunicador). Las variables con mayor asociación se identificaron mediante coeficientes de Pearson (umbral ≥ 0.06) y se visualizaron mediante boxplots, mapas de calor y diagramas de dispersión. Este análisis permitió seleccionar las métricas más informativas para el entrenamiento de modelos predictivos en el capítulo siguiente.

Gráfico, Gráfico de rectángulos

El contenido generado por IA puede ser incorrecto.

#### 3.4.2 Variables agregadas y derivadas a nivel de vídeo

A partir del dataset de segmentos previamente limpiado e imputado (df\_seg\_win\_med), se construyeron variables agregadas por vídeo, con el objetivo de representar de forma resumida el comportamiento multimodal del orador a lo largo del discurso.

Se calcularon estadísticas como la **media, varianza, máximo y mínimo** de variables acústicas (energía, tono, ritmo), visuales (gestos, expresiones, posturas) y emocionales. Para la variable emocion\_audio, que identifica la emoción predominante en cada segmento, se aplicó una codificación **one-hot encoding**, generando variables binarias por cada clase emocional (alegría, tristeza, miedo, etc.).

Estas métricas permitieron caracterizar cuantitativamente cada discurso. Además, se generaron variables derivadas mediante combinaciones entre energía, tono y gestos, con el objetivo de capturar **relaciones expresivas complejas**, como entusiasmo, tensión o implicación corporal.

También se diseñó un **índice de expresividad global**, calculado como una combinación ponderada de movimientos de cabeza (gestualidad) y activación facial (sonrisa, ceño, ojos, etc.).

Finalmente, **se integraron todas las variables generadas en un único dataset final a nivel de vídeo**, almacenado como df\_model.csv, que será utilizado como entrada para el entrenamiento de modelos predictivos en el siguiente capítulo.

## Capítulo 4 – Modelado

Este capítulo presenta el proceso de entrenamiento, evaluación e interpretación del modelo predictivo diseñado para estimar la calidad comunicativa de los discursos analizados. A partir del dataset unificado construido en el capítulo anterior, se entrenó un modelo de clasificación binaria para distinguir entre oradores de alta y baja calidad, utilizando variables multimodales que combinan información textual, acústica, visual y emocional.

Todas las operaciones descritas en este capítulo fueron implementadas en Python y se encuentran documentadas en el **Anexo 5: 5\_Modelo.ipynb**.

### 4.1 Entrenamiento del modelo predictivo

Para el entrenamiento se utilizó el dataset df\_model.csv, con una fila por vídeo y variables que resumen el comportamiento multimodal del orador. El objetivo del modelo es predecir la variable tipo\_comunicador (0 = baja calidad, 1 = alta calidad), y utilizar la probabilidad de pertenencia a la clase 1 como una **medida continua de calidad comunicativa**.

Se seleccionó un modelo **Random Forest** por su equilibrio entre rendimiento y explicabilidad. Sus principales ventajas son:

* **Robustez ante ruido y multicolinealidad** entre variables de naturaleza heterogénea.
* **Capacidad de manejar variables numéricas y categóricas sin necesidad de normalización** previa.
* **Facilidad para interpretar la importancia de las variables**, aspecto clave en este trabajo para proporcionar retroalimentación personalizada al usuario.

La salida probabilística del modelo se interpreta como un **índice continuo de calidad comunicativa**, permitiendo no solo clasificar, sino también ordenar y comparar discursos entre sí.

**4.2 Evaluación del modelo (ampliación)**

Durante el proceso de entrenamiento, se realizó un análisis gráfico de la **precisión (accuracy)** en función del **número de variables seleccionadas**.

Gráfico

El contenido generado por IA puede ser incorrecto.

Se observó que, tras cierto punto, añadir más variables no mejoraba el rendimiento, lo que motivó la búsqueda de una configuración óptima que equilibrara rendimiento y explicabilidad.

Tras diversas pruebas, se seleccionó un conjunto de variables que **mejoran el rendimiento del modelo** y, al mismo tiempo, **facilitan la generación de feedback explicativo para el usuario**, que es uno de los objetivos clave de este trabajo. También se evaluaron varias transformaciones y configuraciones de hiperparámetros, sin mejora significativa. Finalmente, se alcanzó un **accuracy del 72%** y un **AUC de 0.76** con un conjunto de características organizadas en cinco grupos funcionales.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

**🔹 Variables seleccionadas finales por grupo**

* **Texto**:  
  claridad\_oraciones, impacto\_palabras\_poder, impacto\_anecdotas\_detectadas, anecdotas, ejemplos

*Permiten dar feedback sobre estructura, claridad narrativa, uso de ejemplos y recursos persuasivos.*

* **Audio – Energía y Ritmo**:  
  rms\_mean\_mean, rms\_mean\_var, rms\_zcr\_var\_ratio, zcr\_mean\_min

*Indican dinamismo vocal, variabilidad en la energía y articulación, asociados a expresividad.*

* **Audio – Emociones**:  
  emocion\_audio\_happy\_mean, fearful\_mean, angry\_mean, surprised\_mean, disgust\_mean, calm\_mean

*Reflejan el tono emocional predominante en el discurso, clave para la conexión emocional.*

* **Postura y Gestos**:  
  yaw\_max, pitch\_min, apertura\_brazos\_var, expresividad\_facial, expresividad\_gestual, rms\_brazos

*Capturan la expresividad física y gestual del orador, indicadores de implicación corporal.*

* **Otros**:  
  duracion\_video, pmm\_mean

*Contextualizan el discurso en términos de longitud y velocidad media del habla.*

**4.3 Pruebas con vídeos de mayor calidad visual**

Con el objetivo de evaluar si una mayor calidad en la señal visual mejora el rendimiento del modelo, se entrenó nuevamente el Random Forest utilizando únicamente los vídeos que presentaban una **alta cobertura facial**. Específicamente, se seleccionaron aquellos vídeos en los que al menos el **40% de los segmentos** incluían detección de rostro (cara\_detectada = True).

Aunque esta reducción disminuyó el tamaño del conjunto de entrenamiento, los resultados obtenidos fueron similares:

* **Accuracy**: en torno al 70%
* **AUC – ROC**: 0.77

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Esto sugiere que el modelo es **robusto frente a cierta pérdida de señal visual**, y que la información textual y acústica también aporta valor suficiente en ausencia de datos gestuales o posturales. No obstante, se confirma que, cuando están disponibles, las variables visuales enriquecen la capacidad predictiva y explicativa del modelo.

Modelo 4 entrenado con videos de más de 30% de cara detectada, frente a los diferentes niveles de calidad de extracción.

Gráfico

El contenido generado por IA puede ser incorrecto.

Modelo 5 entrenado con videos de más de 40% de cara detectada, frente a los diferentes niveles de calidad de extracción.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Me decanto por el modelo 4

Presenta mayor rendimiento en los conjuntos de datos con valores de cobertura facial entre 4, 5 y 6 (es decir, entre el 40 % y el 70 % de segmentos con rostro detectado), que es precisamente donde estimamos que se concentrará la mayoría de los vídeos analizados cuando el sistema esté en uso real.

Además, muestra un rendimiento excelente en vídeos con **alta calidad visual**, lo que demuestra su capacidad para aprovechar eficazmente la información multimodal disponible (audio, texto y gestos).

Por el contrario, los vídeos con bajos niveles de detección facial suelen corresponder a grabaciones con problemas de encuadre, baja resolución, iluminación deficiente o escasa presencia del orador en primer plano. Si bien estos vídeos se han mantenido durante la fase de entrenamiento —debido al elevado coste de filtrarlos manualmente—, no se espera que este tipo de contenido tenga un peso significativo en contextos reales de uso, donde los usuarios suelen grabar sus propias presentaciones de forma frontal y clara.

En cualquier caso, se ha optado por **guardar ambos modelos entrenados**, lo que permitirá, en producción, **seleccionar dinámicamente el modelo más adecuado** en función de la calidad visual del vídeo introducido por el usuario.

#### 4.4 Explicabilidad del modelo y generación de feedback personalizado

Uno de los objetivos clave de este trabajo no era solo desarrollar un modelo con buen rendimiento predictivo, sino también que dicho modelo fuera **explicable y útil como herramienta de mejora comunicativa**. En otras palabras, se buscaba que el sistema no solo clasificara discursos, sino que pudiera **identificar qué aspectos específicos influyen en la calidad percibida** y ofrecer **retroalimentación comprensible al usuario**.

**Técnicas de explicabilidad: SHAP**

Para interpretar las decisiones del modelo Random Forest, se utilizó la técnica SHAP (SHapley Additive exPlanations), ampliamente reconocida por su capacidad para:

* Asignar un peso cuantitativo a cada variable en cada predicción individual.
* Visualizar la contribución positiva o negativa de cada característica.
* Detectar patrones útiles a nivel global y por grupo de variables.

Esto permite analizar no solo **qué predice el modelo**, sino también **por qué lo hace**, lo cual es esencial para una herramienta centrada en la autoevaluación y el aprendizaje.

**Organización del feedback**

Para facilitar la comprensión y la acción por parte del usuario, las variables utilizadas por el modelo fueron agrupadas temáticamente en cinco bloques funcionales:

1. **Texto**: claridad estructural, uso de palabras de impacto, presencia de anécdotas o ejemplos.  
   *→ Feedback sobre el contenido verbal del discurso.*
2. **Audio – Energía y ritmo**: variabilidad de la energía (RMS), articulación (ZCR), ritmo y pausas.  
   *→ Feedback sobre expresividad vocal y ritmo comunicativo.*
3. **Audio – Emociones**: emociones predominantes estimadas a partir del audio.  
   *→ Feedback sobre el tono emocional general y posibles momentos clave.*
4. **Postura y gestos**: inclinaciones de cabeza, gesticulación (apertura de brazos), expresividad facial.  
   *→ Feedback sobre lenguaje corporal y conexión visual.*
5. **Otros**: duración del vídeo, velocidad media del habla.  
   *→ Indicadores contextuales relevantes para interpretar el resto de variables.*

**Generación del feedback**

A partir de los valores SHAP obtenidos para una predicción, se genera **una explicación textual personalizada** para cada grupo de variables, incluyendo:

* Si el valor está por encima o por debajo del promedio.
* Si su contribución fue positiva o negativa en la predicción.
* Una descripción concreta de qué significa ese patrón y cómo podría mejorarse.

Esto permite ofrecer **recomendaciones prácticas** como:  
*“Tu discurso fue claro y estructurado, pero el uso de ejemplos fue escaso”*, o  
*“El nivel de energía vocal fue bajo, lo que puede haber afectado a la percepción de entusiasmo”*.

**Aplicaciones prácticas**

Esta capacidad de explicar cada decisión modelo convierte la herramienta en un **asistente de evaluación y mejora**, capaz de:

* Guiar a los usuarios en la revisión de sus discursos.
* Sugerir aspectos específicos a trabajar.
* Proporcionar una valoración continua (probabilidad tipo 1) que permite seguir la evolución con el tiempo.

**4.5 Conclusiones del modelado y paso a producción**

El proceso de modelado ha permitido desarrollar un sistema predictivo robusto, explicable y aplicable a contextos reales de evaluación de discursos. A lo largo del capítulo se han construido y evaluado distintas versiones del modelo, optimizando no solo el rendimiento, sino también la capacidad de proporcionar **feedback útil y personalizado**.

Entre las principales conclusiones del modelado destacan:

* La **selección del modelo Random Forest** respondió a su equilibrio entre rendimiento, interpretabilidad y estabilidad frente a datos heterogéneos.
* Se alcanzó un rendimiento estable con una **accuracy del 72 % y un AUC de 0.76**, empleando un conjunto optimizado de variables agrupadas por áreas funcionales (texto, audio, emociones, postura y contexto).
* Se realizaron pruebas sobre subconjuntos de mayor calidad visual, confirmando que, aunque los datos visuales enriquecen el análisis, el modelo mantiene un rendimiento aceptable incluso con menor cobertura facial.
* Gracias a técnicas de explicabilidad como **SHAP**, el modelo permite interpretar cada predicción, identificar qué aspectos influyen más en la calidad comunicativa percibida y generar recomendaciones prácticas.
* Se definió una estrategia de **feedback multimodal por grupos de variables**, que convierte el modelo en una herramienta útil de autoevaluación y mejora de la oratoria.

Con estos elementos, el sistema está preparado para una futura implementación en entornos reales. El modelo final, junto con el pipeline de extracción de características y la generación de feedback, puede ser aplicado a nuevos vídeos para evaluar la calidad comunicativa de forma automática, explicable y accionable.

## Capítulo 5. Prueba de producción

Para validar el sistema en un entorno funcional, se desarrolló un notebook en Google Colab (**Anexo 6 – 6\_produccion.ipynb**) que permite analizar vídeos individuales de forma autónoma. Este cuaderno utiliza funciones definidas en el módulo auxiliar **utils.py (Anexo 7)** y permite procesar vídeos tanto desde una **URL** (por ejemplo, de YouTube) como desde el **almacenamiento local del usuario**.

El objetivo de esta prueba es facilitar el uso práctico del sistema y comprobar su capacidad para **extraer características**, **aplicar el modelo entrenado** y **generar retroalimentación explicativa**, sin necesidad de intervención manual.

A continuación, se presenta un **ejemplo de salida** generado tras procesar un vídeo real.

url: <https://www.ted.com/talks/cameron_russell_looks_aren_t_everything_believe_me_i_m_a_model>

🎯 Nivel de comunicación según el análisis realizado: 0.707

🟢 Nivel de comunicador \*\*medio-alto\*\*

🔷 FEEDBACK GLOBAL

🟢 En general, las características del orador favorecen positivamente la predicción del modelo.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

🟩 FEEDBACK — Grupo: Texto

🔵 El grupo \*\*Texto\*\* tiene una influencia positiva destacada en la predicción.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

🔹 claridad\_oraciones: Claridad sintáctica y gramatical del texto. Frases bien estructuradas facilitan la comprensión.

Tiene un valor de `92.00` frente a un ideal de `139.04` (diferencia `-47.04`). SHAP=+0.004. A pesar de usar estructuras menos claras, estás comunicando eficazmente.

🔹 impacto\_palabras\_poder: Uso de palabras con carga persuasiva, emocional o enfática.

Tiene un valor de `8.00` frente a un ideal de `12.31` (diferencia `-4.31`). SHAP=+0.006. Usas pocas palabras poderosas, pero en tu caso eso mejora la naturalidad.

🔹 impacto\_anecdotas\_detectadas: Impacto estimado de las anécdotas utilizadas en el discurso.

Tiene un valor de `106.00` frente a un ideal de `85.89` (diferencia `+20.11`). SHAP=+0.024. Tus anécdotas están generando un impacto muy positivo en la audiencia.

🔹 anecdotas: Número o presencia de anécdotas personales o ejemplificadoras.

Tiene un valor de `17.00` frente a un ideal de `13.18` (diferencia `+3.82`). SHAP=+0.034. Incluir anécdotas mejora tu cercanía con la audiencia.

🔹 ejemplos: Cantidad de ejemplos concretos usados para explicar ideas.

Tiene un valor de `12.00` frente a un ideal de `11.56` (diferencia `+0.44`). SHAP=+0.006. Los ejemplos que usas están reforzando tu mensaje.

🟩 FEEDBACK — Grupo: Audio - Energía/Ritmo

🔵 El grupo \*\*Audio - Energía/Ritmo\*\* tiene una influencia positiva destacada en la predicción.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

🔹 rms\_mean\_mean: Energía vocal media. Relacionada con el volumen general.

Tiene un valor de `0.05` frente a un ideal de `0.08` (diferencia `-0.02`). SHAP=+0.009. Tu energía vocal está en el rango óptimo.

🔹 rms\_mean\_var: Variabilidad de energía vocal. Indica cambios de volumen que aportan expresividad.

Tiene un valor de `0.00` frente a un ideal de `0.00` (diferencia `-0.00`). SHAP=+0.011. Tu variación de energía está equilibrada.

🔹 rms\_zcr\_var\_ratio: Relación entre variabilidad de energía y ritmo. Indicador de control vocal.

Tiene un valor de `0.55` frente a un ideal de `0.64` (diferencia `-0.09`). SHAP=+0.018. Aunque tu ritmo es más pausado, está ayudando a la claridad.

🔹 zcr\_mean\_min: Duración de las pausas prolongadas en el discurso.

Tiene un valor de `0.08` frente a un ideal de `0.06` (diferencia `+0.02`). SHAP=+0.016. El uso de pausas está equilibrado.

🟩 FEEDBACK — Grupo: Audio - Emociones

🔵 El grupo \*\*Audio - Emociones\*\* tiene una influencia positiva destacada en la predicción.

Gráfico, Gráfico en cascada

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

🔹 emocion\_audio\_happy\_mean: Promedio de emoción de felicidad detectada en el audio. Tiene un valor de `0.10` frente a un ideal de `0.14` (diferencia `-0.05`). SHAP=+0.005. Transmitiste un nivel adecuado de entusiasmo.

🔹 emocion\_audio\_fearful\_mean: Promedio de emoción de miedo detectado en el audio.Tiene un valor de `0.69` frente a un ideal de `0.59` (diferencia `+0.10`). SHAP=+0.017. Una ligera tensión añade humanidad al mensaje.

🔹 emocion\_audio\_angry\_mean: Promedio de emoción de enojo detectado en el audio. Tiene un valor de `0.12` frente a un ideal de `0.11` (diferencia `+0.01`). SHAP=+0.022. Transmitiste una intensidad emocional adecuada.

🔹 emocion\_audio\_surprised\_mean: Promedio de emoción de sorpresa detectada en el audio. Tiene un valor de `0.06` frente a un ideal de `0.06` (diferencia `+0.00`). SHAP=+0.004. El nivel de expresividad está bien regulado.

🔹 emocion\_audio\_disgust\_mean: Promedio de emoción de desagrado detectado en el audio. Tiene un valor de `0.02` frente a un ideal de `0.02` (diferencia `+0.00`). SHAP=+0.016. Transmitiste lo justo de crítica o rechazo emocional.

🔹 emocion\_audio\_calm\_mean: Promedio de emoción de calma detectado en el audio. Tiene un valor de `0.01` frente a un ideal de `0.07` (diferencia `-0.06`). SHAP=+0.007. Tu energía transmite motivación.

🟩 FEEDBACK — Grupo: Postura/Gestos

⚪ El grupo \*\*Postura/Gestos\*\* tiene un impacto neutro o leve en la predicción.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

🔹 yaw\_max: Movimiento máximo de la cabeza (horizontal). Tiene un valor de `73.61` frente a un ideal de `67.15` (diferencia `+6.46`). SHAP=+0.025. Tu movimiento de cabeza aporta naturalidad y conexión.

🔹 pitch\_min: Inclinación vertical mínima de la cabeza (pitch). Indica cuánto baja la cabeza durante el segmento. Tiene un valor de `24.56` frente a un ideal de `41.71` (diferencia `-17.15`). SHAP=-0.004. Poca variación en la inclinación puede hacer que tu expresividad corporal sea limitada.

🔹 apertura\_brazos\_var: Variación en la apertura de brazos. Relacionada con gestualidad abierta. Tiene un valor de `721.46` frente a un ideal de `1355.42` (diferencia `-633.96`). SHAP=-0.016. Podrías abrir más tus gestos para conectar mejor con el público.

🔹 expresividad\_facial: Nivel de expresividad en el rostro. Tiene un valor de `20.15` frente a un ideal de `28.51` (diferencia `-8.36`). SHAP=+0.006. Tu expresión neutra ayuda a enfocarse en el contenido.

🔹 expresividad\_gestual: Expresividad corporal general. Tiene un valor de `87.06` frente a un ideal de `88.72` (diferencia `-1.66`). SHAP=+0.007. Tu sobriedad gestual da serenidad.

🔹 rms\_brazos: Coordinación entre voz y movimiento de brazos. Tiene un valor de `2.15` frente a un ideal de `3.35` (diferencia `-1.20`). SHAP=-0.026. Más gestos coordinados con tu voz pueden hacerte más persuasivo.

🟩 FEEDBACK — Grupo: otros

⚪ El grupo \*\*otros\*\* tiene un impacto neutro o leve en la predicción.

Gráfico

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

🔹 duracion\_video: Duración total del video analizado. Tiene un valor de `561.02` frente a un ideal de `866.82` (diferencia `-305.80`). SHAP=-0.011. Podrías extender un poco más el contenido para reforzar tus ideas.

🔹 pmm\_mean: Velocidad de habla (palabras por minuto). Tiene un valor de `206.33` frente a un ideal de `193.62` (diferencia `+12.72`). SHAP=+0.021. Tu velocidad transmite dinamismo y energía.

## Capítulo 6 – Conclusiones y líneas futuras

### 6.1 Conclusiones generales

Este trabajo ha demostrado la viabilidad técnica y práctica de construir un sistema automático capaz de analizar discursos en vídeo mediante un enfoque multimodal, combinando texto, audio y vídeo. El objetivo principal ha sido identificar **características asociadas a una comunicación efectiva** y ofrecer al usuario un **feedback explicable y accionable** sobre su desempeño.

A través de una arquitectura modular, un pipeline optimizado y un modelo explicable basado en Random Forest, el sistema ha alcanzado un rendimiento competitivo (72 % de accuracy, AUC de 0.76), siendo capaz de identificar patrones útiles en presentaciones grabadas con características variables.

Más allá del rendimiento, el valor diferencial del sistema reside en su **capacidad de ofrecer retroalimentación personalizada**, basada en las dimensiones clave de la comunicación (estructura verbal, expresividad vocal, emocionalidad, lenguaje corporal). Esta funcionalidad convierte el modelo en una **herramienta útil para el aprendizaje y la mejora continua**, tanto a nivel individual como organizacional.

### 6.2 Valor de negocio y posibles aplicaciones

Desde un punto de vista de negocio, este sistema tiene un alto potencial de aplicación en contextos como:

**🧑‍🏫 Formación y educación**

* Uso en cursos de oratoria, presentaciones, debate o comunicación efectiva.
* Evaluación automática de presentaciones en entornos académicos.
* Retroalimentación personalizada para estudiantes en programas online o híbridos.

**💼 Recursos humanos y selección**

* Evaluación de competencias comunicativas en entrevistas grabadas o vídeos de presentación.
* Detección de habilidades blandas (soft skills) relevantes para puestos de liderazgo o ventas.

**🧑‍⚕️ Coaching y desarrollo personal**

* Herramienta de apoyo para coaches y entrenadores de comunicación.
* Seguimiento del progreso en habilidades expresivas con indicadores objetivos.

**📱 Integración en plataformas digitales**

* Como componente embebido en plataformas de e-learning, videollamadas o grabación de pitches.
* Potencial integración en apps de autoevaluación para profesionales, docentes o creadores de contenido.

### 6.3 Ventajas competitivas

* 🔍 **Explicabilidad**: El sistema no solo clasifica, sino que **explica** cada evaluación, indicando qué mejorar.
* 🤖 **Automatización total**: Desde el vídeo original hasta el feedback final, sin intervención humana.
* 📊 **Medición objetiva**: Basado en modelos y características observables, sin sesgos subjetivos.
* 🔄 **Escalabilidad**: Puede adaptarse a distintos idiomas, tipos de discurso y niveles de exigencia.

### 6.4 Limitaciones y líneas futuras

* La calidad de los resultados depende en parte de la calidad visual del vídeo (presencia de rostro, iluminación).
* Algunas características gestuales pueden perderse si el orador no está centrado o bien encuadrado.
* El análisis emocional podría mejorarse con modelos más avanzados o datos contextuales (tema, audiencia).

Líneas de mejora propuestas:

* Entrenar modelos más especializados por tipo de discurso (académico, comercial, divulgativo).
* Integrar un LLM (p. ej., GPT) para transformar las señales estructuradas (métricas + valores SHAP) en retroalimentación natural, empática y contextualizada
* Incorporar análisis del entorno visual (diapositivas, fondo, interacción con la audiencia).
* Mejorar la visualización del feedback con dashboards interactivos o informes descargables.
* Evolucionar desde una valoración agregada por vídeo a un feedback temporalmente localizado, señalando instantes o tramos concretos donde actuar.
* Validar el sistema con usuarios reales en contextos formativos o profesionales.

### 6.5 Conclusión final.

Este trabajo no solo demuestra que es posible analizar la calidad comunicativa de discursos de forma automatizada, sino que abre la puerta a sistemas inteligentes que **no se limitan a evaluar, sino que ayudan a mejorar**.

Desde una perspectiva de negocio, se trata de una tecnología con **alto valor añadido, replicable y adaptable**, que puede ser integrada en soluciones educativas, profesionales o corporativas, con un impacto directo en el desarrollo de habilidades comunicativas, uno de los activos más valorados en el mundo actual.