Memoria Técnica: Sistema de Detección de Emociones en Audio (SER)

Procesamiento del habla, visión e interacción multimodal

Autores:

Daniel Marchena Jiménez
Javier Arias Fuentes
Zohair Mouhim Gharafi

Índice

1. Introducción y Contexto

El objetivo de este proyecto es desarrollar un sistema de Inteligencia Artificial capaz de clasificar el estado emocional de un hablante basándose exclusivamente en las características acústicas de su voz. Siguiendo los requisitos establecidos en la "Práctica 3", se ha prescindido de cualquier técnica de Transcripción Automática del Habla (ASR) o Procesamiento de Lenguaje Natural (NLP), centrando el análisis en los componentes paralingüísticos (cómo se dice) y no en el contenido lingüístico (qué se dice).

El sistema se ha diseñado para clasificar audios en categorías emocionales simplificadas (Positiva, Neutra, Negativa) o extendidas (Enfadado, Feliz, Neutro, Triste), utilizando un enfoque híbrido: extracción de características mediante Python (librosa) y modelado mediante minería de datos visual en Orange Data Mining.

2. Selección del Dataset y Preprocesamiento de Audio

2.1. Elección del Corpus de Datos

1. **Dataset de entrenamiento:** Para el entrenamiento del modelo, se ha seleccionado el dataset `stapesai/ssi-speech-emotion-recognition`. La elección de este conjunto de datos se justifica por los siguientes factores:
    a. **Consistencia y Accesibilidad:** Al utilizar un dataset curado y alojado en Hugging Face, se garantiza el acceso a datos estructurados y listos para su procesamiento, evitando inconsistencias de formato comunes en la recolección manual de archivos de audio.
    b. **Calidad de Audio:** Este dataset proporciona grabaciones con una relación señal-ruido (SNR) adecuada, lo que facilita la extracción de características limpias (MFCCs, Chroma) sin la interferencia excesiva de ruido ambiental.
    c. **Etiquetado Fiable:** Las muestras contienen etiquetas emocionales claras, lo cual es ideal para el entrenamiento supervisado de modelos base, permitiendo al sistema aprender patrones acústicos prototípicos de cada emoción.
2. **Dataset de Control:** `Ryerson Audio-Visual Database of Emotional Speech and Song` o `RAVDESS`. Al ser grabaciones de actores profesionales en estudio anecoico, nos sirve como "Ground Truth" de emociones arquetípicas.
3. **Dataset Experimental:** Grabaciones realizadas por el equipo para testear el sistema ante micrófonos no profesionales y ruido ambiente.

2.2. Extracción de Características (Script Python)

Parámetro	Detalle
Herramienta	Python + Librería `librosa`
Repositorio	• `ssi-dataset-cc-extractor` • `RAVDESS`
Decisión de Diseño	Script externo para control total sobre ventana de análisis y parámetros matemáticos. El audio no estructurado se convierte a datos tabulares para ML.

Características Extraídas y Justificación:
Se extrajeron las características físicas del archivo `.wav` y se calcularon sus medias para obtener un vector de longitud fija por cada audio. Las principales variables seleccionadas fueron:

- **MFCCs (Mel-frequency cepstral coefficients):** Se extrajeron 40 coeficientes.
  - *Por qué:* Los MFCC replican la audición humana y capturan el timbre de la voz. Son la característica más discriminante en el reconocimiento de emociones, ya que permiten diferenciar la "aspereza" de la ira frente a la "suavidad" de la calma, independientemente de lo que se diga.
- **Chroma (Cromagrama):**
  - *Por qué:* Relacionado con la armonía y el tono musical. Ayuda a distinguir variaciones tonales intensas.
- **Mel Spectrogram & Spectral Contrast:**
  - *Por qué:* Miden la distribución de energía en diferentes frecuencias y la diferencia entre picos y valles del espectro, útiles para detectar la intensidad emocional (Arousal).

El resultado de este proceso fue un archivo estructurado (Excel/CSV) donde cada fila es un audio y cada columna una característica numérica, listo para Orange.

3. Metodología en Orange Data Mining

El flujo de trabajo (`Practica_3.ows`) se ha diseñado para cargar, procesar, entrenar y validar los modelos. A continuación se detalla la función de cada nodo y la decisión detrás de su uso:

3.1. Estrategias de Experimentación

Hemos implementado dos tuberías de procesamiento distintas (archivos .ows) para evaluar cómo la definición de las clases afecta al rendimiento:

- **Estrategia A:** Agrupación en 3 clases de emociones. Se han fusionado emociones semánticamente cercanas en tres emociones principales (Positiva, negativa, neutra) para reducir la complejidad y buscar una mayor tasa de acierto global. Dado que esta fusión generó una "super-clase" con muchas más muestras que el resto, se produjo un **desbalance de clases**. Para evitar que el modelo adquiriera un **sesgo** hacia la emoción predominante, aplicamos **submuestreo**, limitando artificialmente la cantidad de instancias de dicha clase para equilibrar la distribución antes del entrenamiento.
- **Estrategia B:** Selección estándar de 4 emociones básicas. Se ha filtrado el dataset para conservar únicamente las emociones universales de Ekman: Ira, Tristeza, Felicidad y Neutral, descartando el resto.

3.2. Carga y Transformación de Datos

- **File:** Importación del dataset generado por el script Python (`ssi_custom_features.xlsx`).
- **Edit Domain:**
  - *Decisión Crítica:* Este nodo se utiliza para cumplir con el requisito de simplificación de emociones. Dado que clasificar 7 emociones distintas es complejo y puede llevar a error, se utilizó este nodo para mapear las clases originales a un subconjunto más robusto (ej. agrupar *Happy* y *Surprise* en "Positivo", y *Anger* y *Sadness* en "Negativo"), o para asegurar que la variable objetivo sea categórica.
- **Select Columns:**
  - Se definió la columna `emotion` (o la variable mapeada) como **Target** y los coeficientes numéricos (MFCCs, Chroma, etc.) como **Features**. Los metadatos como el nombre del archivo se excluyeron del entrenamiento para evitar sesgos.

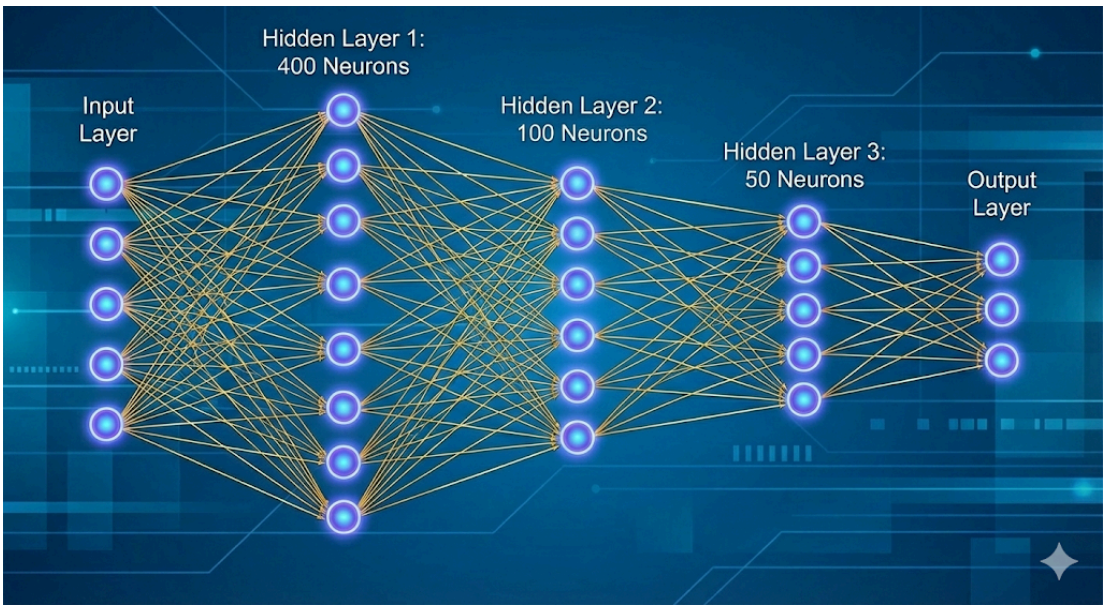
3.3. Selección de Características

- **Rank (Ranking):**
    - *Por qué:* El script extrae muchas columnas (más de 40). No todas aportan información útil; algunas pueden introducir ruido.
    - *Método:* Se utilizó **Information Gain (Ganancia de Información)** o **ReliefF** para ordenar qué características físicas discriminan mejor entre emociones. Esto nos permite ver, por ejemplo, que los primeros coeficientes MFCC suelen ser más importantes que los últimos.
    - *Conclusión*: Al realizar pruebas excluyendo las columnas menos significativas, observamos que la precisión del modelo **empeoraba**. Esto confirma que, para nuestro volumen de datos, utilizar las 40 variables no supone un problema. Por tanto, decidimos mantenerlas todas para no perder información útil del audio, descartando que esto fuera a causar problemas de aprendizaje o sobreajuste.

3.4. Configuración de Red Neuronal

Hemos configurado la red neuronal por defecto añadiendo 2 capas ocultas extra y configurando el número de neuronas en forma de embudo.

- **Arquitectura:** [400, 100, 50] neuronas.



- **Justificación del Diseño:**
    - Capa 1 (400 neuronas): Una capa inicial ancha para capturar una gran variedad de combinaciones de bajo nivel de los 40 coeficientes MFCC de entrada.
    - Capa 2 (100 neuronas) y Capa 3 (50 neuronas): Reducción progresiva de la dimensionalidad. Esta estructura obliga a la red a "comprimir" la información, sintetizando los patrones más relevantes y descartando el ruido acústico.

- **Función de Activación:** ReLu (Rectified Linear Unit), seleccionada por su eficiencia computacional y por mitigar el problema del desvanecimiento del gradiente en redes de varias capas.
- **Solver:** Adam, optimizado para manejar grandes volúmenes de datos y converger más rápido que el descenso de gradiente estocástico tradicional.

3.5. Configuración del Flujo de Validación

El diseño en Orange combina dos métodos de evaluación simultáneos:

1. **Validación Interna (Test & Score):** Se aplica *Cross-Validation (k-fold)* sobre el dataset de entrenamiento. Esto elimina el sesgo de partición y verifica que el modelo es estadísticamente estable con los datos conocidos.
2. **Inferencia Externa (Predictions):** Se entrena el modelo con la totalidad de los datos de entrenamiento y se lanzan predicciones sobre los datasets de `RAVDESS` y `Voces Propias`. Esto mide el rendimiento real ante el cambio de micrófono y entorno.

4. Modelado y Algoritmos Seleccionados

Se han comparado dos familias de algoritmos muy distintas para analizar cuál se adapta mejor a las características espectrales:

4.1. Random Forest

Configuración	**Justificación**
Conjunto de árboles de decisión (ej. 10 a 100 árboles).	Es el algoritmo "estándar de oro" para datos tabulares.
Ventajas	Funciona excepcionalmente bien con MFCCs, es robusto frente al ruido y no requiere que los datos estén normalizados (a escala). Maneja bien la alta dimensionalidad de los 40 coeficientes.

4.2. Neural Network (Red Neuronal - MLP)

Configuración	**Justificación**
Perceptrón Multicapa con capas ocultas (ej. una capa de 100 neuronas, activación ReLu).	Es capaz de capturar relaciones no lineales complejas entre las frecuencias de audio.
Consideraciones	Requiere más datos para converger y es más sensible a la falta de normalización que los árboles.

5. Análisis de Resultados

5.1. Evaluación (Confusion Matrix y ROC)

Los nodos **Confusion Matrix** y **ROC Analysis** se utilizaron para interpretar el rendimiento:

Herramienta	**Objetivo**	**Observación Típica**
Matriz de Confusión	Ver no solo cuánto acierta el modelo, sino cómo se equivoca.	Es común que el modelo confunda emociones de alta energía (Angry vs Happy) o de baja energía (Sad vs Neutral).
Curva ROC	Mostrar la capacidad del modelo para separar las clases.	Un área bajo la curva (AUC) superior a 0.8 indica un buen modelo.

5.2. Comparativa

En general, Random Forest suele ofrecer un rendimiento más estable e inmediato con datasets de tamaño medio (< 5000 audios) y características MFCC, gracias a su robustez intrínseca. Por otro lado, la Red Neuronal, aunque sensible al *overfitting* con datasets pequeños, tiene un mayor potencial para capturar patrones más finos y complejos en grandes volúmenes de datos.

Modelo	**Ventaja Principal con SER**	**Rendimiento Típico (MFCC)**
Random Forest	Estabilidad, robustez, no requiere normalización.	Alto rendimiento con datasets medianos.
Neural Network (MLP)	Capacidad para capturar relaciones no lineales complejas.	Mayor potencial con datasets grandes; sensible a *overfitting*.

5.3. Problema de entrenamiento con RAVDESS

Se ha observado un fenómeno contraintuitivo: los modelos obtienen un rendimiento notablemente superior **(88-94% de precisión)** en el dataset externo `RAVDESS` que en el propio dataset de entrenamiento (~70%).

Interpretación Técnica:

- Este resultado sugiere que el modelo ha aprendido a identificar arquetipos emocionales de alta intensidad. `RAVDESS`, al ser interpretado por actores, presenta emociones "de caricatura" con rasgos acústicos muy marcados y separables. Por el contrario, las grabaciones naturales o menos actuadas presentan fronteras difusas entre emociones (sutileza), lo que dificulta la clasificación.

5.4. Problemas en Producción

El rendimiento desciende en el dataset de `Voces Propias`. Esto evidencia la dependencia del modelo respecto a la calidad del canal de audio (micrófono y ruido de fondo) y la dificultad de simular emociones genuinas sin entrenamiento actoral. El modelo busca patrones de intensidad y timbre que una persona no entrenada no siempre produce al fingir una emoción.

6. Conclusión

El desarrollo de este sistema ha demostrado que es posible identificar emociones sin analizar el texto, basándose puramente en la física del sonido. La combinación de la extracción precisa de características con `librosa` y la validación rápida de modelos en Orange ha permitido iterar ágilmente en el diseño del sistema.

La decisión más impactante fue el uso de MFCCs, que demostraron ser los predictores más fuertes. Como mejora futura, se podría implementar una red neuronal convolucional (CNN) directamente sobre los espectrogramas, eliminando la necesidad de calcular medias estadísticas, aunque esto requeriría mayor potencia computacional.