ENFOQUE DUAL PARA LA DETECCIÓN DE EMOCIONES CRÍTICAS EN EL HABLA

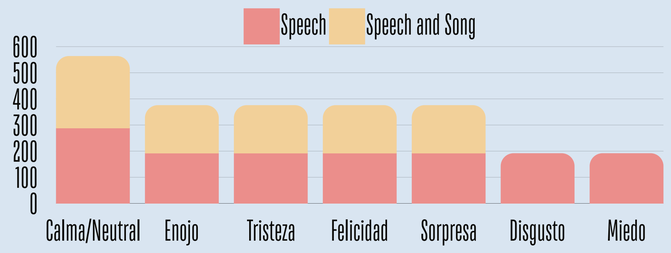
DELFINA CHAVEZ BLASI Y JULIETA GARCIA PEREYRA
dchavezblasi@udesa.edu.ar, jgarciapereyra@udesa.edu.ar
Ingeniería en Inteligencia Artificial, Universidad de San Andrés, Victoria, Buenos Aires, Argentina

INTRODUCCIÓN

Las emociones son fundamentales para el desarrollo y funcionamiento integral de las personas. La **detección de emociones** a partir del habla (SER) tiene diversas aplicaciones, como la asistencia en salud mental y la mejora de la interacción humano-computadora. Sin embargo, muchas veces resulta de especial interés identificar además la **intensidad de una emoción**; dada la complejidad del comportamiento humano, es necesario prestar especial atención a emociones que pueden representar un riesgo potencial, como la tristeza, el enojo y el miedo en niveles extremos. Para abordar esta necesidad, desarrollamos un **modelo dual** con un primer clasificador que predice una emoción a partir del habla y, posteriormente, si la emoción predicha tiene el potencial de ser peligrosa ("crítica"), se envía a un segundo modelo que realiza una clasificación binaria para identificar si la intensidad de la emoción es normal o fuerte.

DATASET

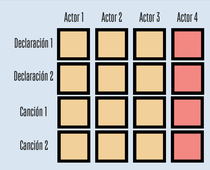
Tanto para el entrenamiento como para la evaluación se utilizaron los audios del **dataset RAVDESS** que cuenta con grabaciones de actores diciendo y cantando dos declaraciones léxicamente coincidentes, con 8 emociones. Para facilitar el aprendizaje de los modelos juntamos la emoción neutral y la de calma y obtuvimos la siguiente distribución de audios:



VALIDACIÓN

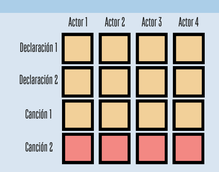
Para realizar la búsqueda de hiperparámetros y evaluar el desempeño de los modelos, al igual que asegurar su capacidad de generalización a datos no vistos (**eliminar correlaciones espurias**), se implementaron tres enfoques de **validación cruzada**: separando en grupos por actor, por declaración o por actor y declaración. Dentro de estos grupos se puede normalizar de dos maneras: una **normalización global** estandarizando el train set y, para fines unicamente experimentales, una **normalización por actor**, estandarizando por feature para cada actor en particular y así eliminar las variaciones individuales de la voz .

Validación por actor



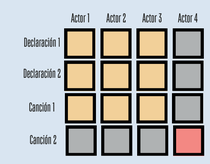
Se asegura que los datos de entrenamiento y validación no compartan actores. Esta garantiza que el modelo pueda generalizar a voces nuevas.

Validación por declaración



El modelo entrena con los datos de una declaración específica y se valida con otra, evaluando la capacidad del modelo para generalizar a diferentes contenidos del habla.

Validación por actor y por declaración



Se elige un actor y una declaración para validación. El reste de los datos correspondientes a ese actor y a esa declaración se descartan. El modelo se entrena con los datos restantes. Este método proporciona una evaluación más rigurosa, descartando cualquier correlación espuria.

FEATURES

Se emplearon características acústicas del conjunto eGeMAPS, extraídas mediante el software openSMILE. Este conjunto incluye **88 características relevantes** para identificar emociones, entre ellas:

- Frecuencia Fundamental (F0): Refleja variaciones en el tono.
- Energía y Amplitud: Indica la intensidad y volumen del habla.
- Espectro de Frecuencia: Captura la calidad del timbre de la voz.
- Medidas de Calidad de la Voz: Miden la estabilidad y variación, como el jitter.

MODELOS

Para el Modelo Dual, se experimentó con modelos como Random Forest, MLP, Gradient Boosting y Regresión Logística. Los modelos finales utilizados fueron los siguientes:

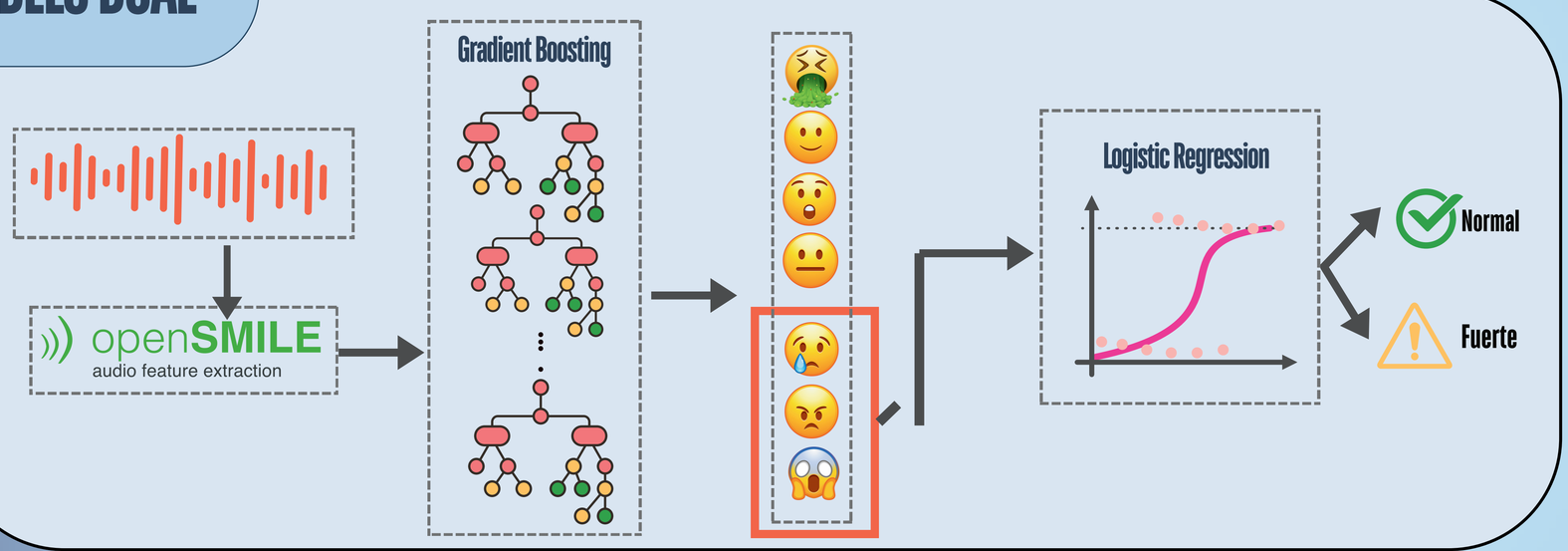
Detección de Emociones : **Gradient Boosting**

- Input: 88 características acústicas de eGeMaps (Geneva Minimalistic Acoustic Parameter Set) sobre un audio en formato .wav
- Output: emoción, etiqueta del 1 - 7 (calma, felicidad, tristeza, enojo, miedo, disgusto o sorpresa)
- Hiperparámetros –> n estimators: 300, learning rate: 0.105, max depth: 5, min samples per leaf: 3.

Detección de Intensidad : **Regresión Logística**

- Este modelo estima la probabilidad condicional: clasifica la intensidad dada la emoción identificada por el modelo previo. Esto aporta información relevante al modelo para predecir con mayor precisión la intensidad emocional ya que la intensidad y la emoción están correlacionadas.
- Input: 88 características acústicas + emoción predicha en el modelo anterior. Esta emoción se acopla a las features mediante vectores one-hot-encode. El primer vector corresponde a la emoción de tristeza, el segundo a enojo y el tercero a miedo.
- Output: etiqueta binaria (1 = fuerte, 0 = normal)
- Hiperparámetros –> C: 2, solver: lbfgs, max iter: 500, class weight: {1:1 , 2 =:2.6}

MODELO DUAL



1. **Obtención de datos**: audios en formato .wav
2. **Procesamiento y armado de dataset**: extracción de 88 features eGeMAPs mediante el modulo de python de openSMILE y organizados en un .csv
3. **Predicción de la emoción**: vector de 88 features enviado al modelo de Gradient Boosting. Clasificación en una de las 7 emociones
4. **Predicción de la intensidad**: en caso de ser una emoción potencialmente crítica (tristeza, enojo, miedo) , una Regresión Logística predice su intensidad (normal o fuerte).

RESULTADOS

Figura I. Matriz de confusión de Regresión Logística sobre set de test

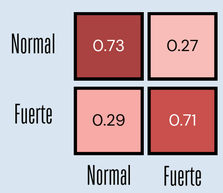
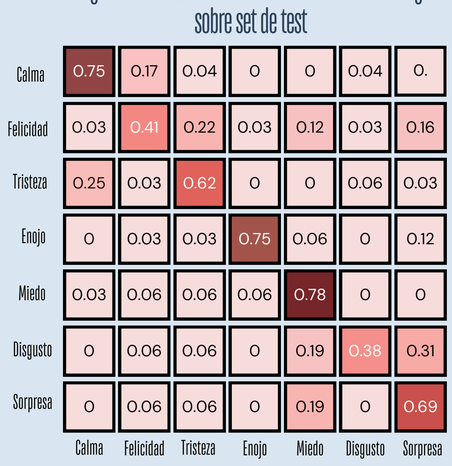


Figura II. Matriz de confusión de Gradient Boosting sobre set de test



En las Figuras I y II se muestran los resultados obtenidos para cada modelo implementado, tuneado con validación cruzada por actor y por discurso y evaluado sobre los datos de test (dos actores no usados para el entrenamiento).

En la predicción de emociones se obtuvo una accuracy general de **0.689** sobre los grupos de validación y **0.649** sobre el set de test. Para el caso de la predicción por intensidad, el mayor desafío fue que el modelo lograra identificar las complejas relaciones entre las características pero que no sobreajuste porque se tenían únicamente 1024 audios para el entrenamiento. A su vez, considerando que el objetivo del proyecto es lograr identificar con la mayor precision posible aquellas emociones cuya intensidad extrema puede presentar un riesgo y requerir de una intervención externa, se le estableció un peso de 1 a la clase 'normal' y de 2.6 a 'fuerte'. Esta ponderación resultó óptima para balancear el trade-off entre minimizar falsos positivos y falsos negativos, asegurando una detección precisa de emociones intensas sin generar demasiadas alertas innecesarias que pueden conllevar un costo asociado. La accuracy del modelo sobre el set de validación fue de **0.85** y sobre el de test final fue de **0.725.** Sin embargo, para el caso del Modelo Dual, como se toma las emociones predichas por el Gradient Boosting y usarlas como entrada para estimar la intensidad, se propaga el error del primer modelo, lo que reduce la accuracy final del Modelo Dual a **0.645.** Otro resultado importante a destacar es que en todos los experimentos la accuracy fue mayor al normalizar por actor que al normalizar los datos globalmente, lo cual indica que individualidad y la forma de expresión emocional de cada actor tienen un impacto significativo en la precisión del modelo.

CONCLUSIÓN

Como conclusiones principales, corroboramos que utilizar **cross-validation agrupando por actor y declaración** fue la mejor manera de tunear hiperparámetros y evaluar la generalización del modelo al eliminar correlaciones espurias asociadas a la voz (actor) y a las palabras (discurso). Adicionalmente, la **normalización por actor** mejoró los resultados al reducir la variabilidad interindividual, aunque no es viable en entornos reales con datos escasos. El **Gradient Boosting** se destacó como el modelo más efectivo para la predicción de emociones, mientras que la **Regresión Logística** fue eficaz en la detección de la intensidad emocional para emociones de riesgo potencial.

Este estudio presenta múltiples oportunidades para trabajos futuros y mejoras, entre las cuales se destacan las siguientes.

- **Ampliación del Dataset:** El datatset RAVDESS contiene grabaciones actuadas, lo que puede introducir sesgos y limitar la aplicabilidad a situaciones reales, donde las emociones son genuinas.
- **Diversidad de Actores:** Considerar la diversidad de actores es fundamental debido a la relevancia de la individualidad y la expresión única de cada actor, como destacaron nuestros resultados.
- **Intensidad Emocional:** Se pueden recolectar datos que abarquen diversos niveles de intensidad emocional para enriquecer el conjunto de datos y mejorar la capacidad predictiva del modelo.
- **Modelos Más Complejos:** Explorar redes neuronales convolucionales (CNN) y recurrentes (RNN) para capturar características temporales y secuenciales con mayor detalle.
- **Combinación de Modalidades:** Integrar características de la voz con expresiones faciales o datos biométricos para aumentar la precisión y robustez del sistema.