uc3m | Universidad Carlos III de Madrid

Grado en Ingeniería Informática
2020-2021

Trabajo Fin de Grado

"Reconocimiento acústico de las emociones del conductor"

Javier Hermida Lario

Tutor/es

Agapito Ismael Ledezma Espino

Leganés, Julio 2021

ABSTRACT

The human factor involved in driving causes between 80% and 90% of traffic accidents that involve massive social and economic losses. The great influence that emotions have in these types of incidents motivates this project, which aims to create a component that identifies the emotions of the people inside a vehicle through speech, and that is created with the purpose of being able to be integrated in a functional ADAS in the future.

This designed component uses audio data processing techniques and Deep Learning to create models capable of recognizing specific patterns related to specific emotions through speech. This problem is tackled from two different data perspectives, producing a final model with a 93% accuracy in test data.

The integration and validation of the designed component is developed in a simulated environment of the real situation, which would have the different components of an ADAS that are not developed in this project.

Key Words: Machine Learning, Deep Learning, Convolutional Neural Networks, MFCC, Emotion Recognition, ADAS, Emotions in driving.

RESUMEN

El factor humano involucrado en la conducción provoca entre el 80% y 90% de los accidentes de tráfico que se producen y que causan grandes pérdidas sociales y económicas. La gran influencia que tienen las emociones humanas en este tipo de sucesos motiva este proyecto, cuyo objetivo es crear una componente encargada de identificar mediante el habla la emoción de los integrantes del vehículo, y que se diseña con la intención de poder ser integrado en un ADAS operativo en un futuro.

La componente diseñada utiliza técnicas de procesado de datos de audio y *Deep Learning* para crear un modelo capaz de reconocer los patrones específicos relacionados con ciertas emociones a través del habla. Este problema es abordado desde dos perspectivas distintas con respecto al formato de datos de entrada del modelo, produciendo un modelo final con un 93% de *accuracy* para el conjunto de test.

Se realiza la integración y validación de la componente diseñada en un contexto simulado de la situación real, la cual constaría del resto de elementos de un ADAS y que no se abordan en este proyecto.

Palabras Clave: Aprendizaje Automático, *Deep Learning*, Redes de Neuronas Convolucionales, MFCC, Reconocimiento de Emociones, ADAS, Emociones en la conducción.

DEDICATORIA

Tras cuatro años de carrera que han pasado más rápido de lo que podría imaginar en un principio, este trabajo pone la guinda del pastel a una etapa increíble.

Muchas gracias al grupo de las Turing Machines por las risas y las noches para el recuerdo y las que quedan por venir.

Especial gracias a mi consultor y proveedor de tortillas de patatas que me ha ayudado tanto en los momentos de crisis.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

1.1. Contexto y motivación

La invención del primer automóvil de combustión de gasolina en 1886 por parte de Karl Friedrich Benz [1] y su posterior producción en masa a manos de Henry Ford [2], ha supuesto un cambio trascendental en la sociedad y economía global.

En 2018, se estimó en más de 500 el ratio de coches por cada 1000 habitantes en la Unión Europea [3]; una cifra que a principios de 2021 aumentó en un 9% y la cuál se prevé que siga en aumento [4].

La introducción del automóvil como elemento básico y habitual en el día a día de las personas, ha incrementado su la calidad de vida, al ofrecer independencia y facilidades de movilidad que previamente eran inalcanzables. Sin embargo, también trajo consigo una media de 1.3 millones de fallecidos anualmente en accidentes de tránsito, siendo la principal causa de muerte en población comprendida entre 15 y 29 años [5]. Una cifra que, mediante numerosas campañas de concienciación, medidas y el aumento de las medidas de seguridad de los vehículos, se ha conseguido reducir un 50% desde 2006 en España como se puede observar en la Fig. 1.1.

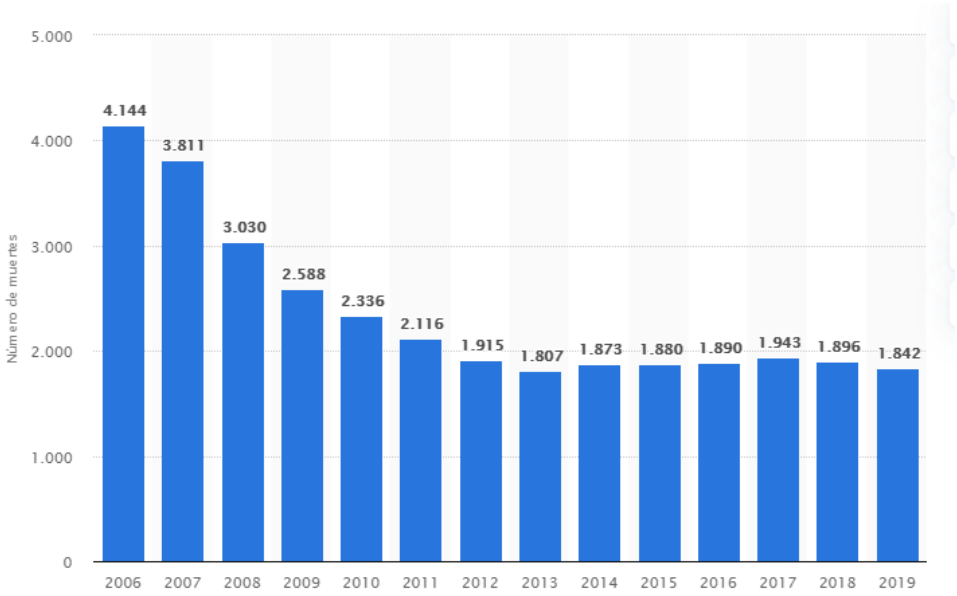


Fig. 1.1. Número de muertes en España entre 2006 y 2019 [6]

Un factor determinante en el descenso de las pérdidas humanas, y por consecuente materiales, es la mejora tecnológica de los automóviles en cuanto a construcción, manejabilidad, sistemas de seguridad, etc. Sin embargo, un factor fundamental que es independiente del propio coche, es el humano que lo conduce. Es por eso por lo que entre el 80% y 90% de los accidentes de tráfico son provocados por fallos humanos como pueden ser distracciones, fatiga, temeridad o consumo de alcohol [7].

Para aplacar este factor humano, la mayoría de las novedades actuales en seguridad vial son introducidas mediante los denominados Sistemas Avanzados de Asistencia a la Conducción (ADAS), cuya función es ayudar al conductor a prevenir accidentes e introducir cierto nivel de automatización a la conducción. Dentro de las funciones que desempeñan ADAS, hay un tipo específico denominadas funciones de confort que se encargan de alertar al conductor de situaciones peligrosas con respecto a la carretera o de su propio estado [8]. Sin embargo, este tipo de ADAS centra su atención principalmente en el estado del conductor, mientras que los pasajeros también influyen en el desempeño de la conducción. Por ejemplo, una discusión en la que el copiloto provoca un efecto negativo en la conducción a pesar de no ser quien la desempeña. Es por eso, que con el fin de reducir los daños humanos y materiales consecuencia de los accidentes de tráfico y mejorar la efectividad de otros ADAS, se propone una posible solución que utiliza técnicas basadas en inteligencia artificial para monitorizar el estado emocional de todos los ocupantes de un vehículo, y poder así utilizar esa información en otros sistemas que tomen las medidas oportunas para evitar o mitigar un accidente.

1.2. Objetivos

Considerando los accidentes de tráfico como una de las principales causas no naturales de fallecimientos y teniendo en cuenta el impacto socio-económico que suponen, la motivación de este proyecto es reducir este tipo de incidentes y así atenuar sus consecuencias negativas.

El objetivo principal del proyecto es diseñar una componente que pueda formar parte de un Sistema Avanzado de Asistencia a la Conducción, la cual se encargue de predecir la emoción del conductor u otro integrante de un vehículo mediante el uso de técnicas de *Deep Learning* y Computación Afectiva.

Para llevar a cabo el objetivo principal es necesario completar una serie de objetivos específicos:

- Estudiar el efecto tanto positivo como negativo de las emociones en las personas y en concreto en la tarea de la conducción
- Estudiar y poner en práctica las distintas técnicas de extracción de características del audio
- Definir un modelo de *Deep Learning* capaz de predecir la emoción de un integrante de un vehículo a través del habla
- Comparar los resultados entre distintas aproximaciones en el tratamiento de los datos

1.3. Marco Regulador

Este apartado tiene como objetivo establecer el contexto legislativo vigente en el periodo en el que se desarrolla este proyecto. Para ello se desarrolla la regulación de uso de tecnología de automatización en automóviles, la legislación acerca de la protección de datos en grabaciones de voz y las licencias de los productos y datos utilizados.

En primer lugar, es necesario establecer la capacidad de actuación que tiene el sistema diseñado en la conducción para así poder establecer la regulación específica para ese nivel de automatización. El sistema en el que se puede integrar la componente diseñada puede tener diversas funciones como pueden ser alertar al conductor o servir para adaptar otros sistemas, pero en ningún caso se diseña para ser integrada en un sistema que influya sobre el control de la dirección del vehículo. Por este motivo, la única legislación que la concierne es la establecida en la Instrucción 15/V-113, expedida por el Ministerio de Interior en conjunto con la Dirección General de Tráfico, en la cual se crea un permiso para la realización de experimentos con vehículos autónomos. De entre los requisitos para obtener este permiso, los cuales tendrían que ser tenidos en cuenta en caso de desplegar este sistema, destacan la necesidad de identificar el vehículo, tener un seguro en regla o la necesidad de acreditar el sistema autónomo mediante autoridades externas.

Otro aspecto que se necesita considerar es la legislación relacionada con el uso de grabaciones de voz. Dado que la Ley Orgánica 3/2018 de Protección de Datos (LOPDGDD) legisla sobre los datos que permiten identificar la identidad de una persona y, dado que los aspectos fisiológicos del habla permiten identificar a una persona, este tipo de dato se encuentra regulado por ella.

Para cumplir las obligaciones establecidas por la LOPDGDD se tendrían que tomar las siguientes medidas en caso de la implementación:

- No obtener grabaciones del usuario hasta que se aporte una autorización expresa del mismo
- No compartir las grabaciones del usuario con terceros sin tener su autorización expresa
- Permitir al usuario revocar la autorización en cualquier momento y la inmediata paralización del uso de sus datos

En último lugar se encuentran las licencias que regulan los recursos externos utilizados. Estas se muestran en las TABLA 1.1 y TABLA 1.2, donde están definidas por su nombre, su descripción y aquellos recursos utilizados que tienen este tipo de licencia.

TABLA 1.1. LICENCIAS UTILIZADAS EN EL PROYECTO (PARTE 1)

Nombre	Descripción	Recursos asociados
Creative Commons Attribution License	Permite distribuir, mezclar, adaptar y construir en un trabajo propio incluso comercial siempre que se cite a los autores	RAVDESS Dataset
Creative Commons license Attribution-NonCommercial-NoDerivatives 4.0 International.	Permite copiar, utilizar y distribuir en un trabajo propio siempre que se cite a los autores, no se utilice para fines comerciales ni se difunda en caso de modificarlo	TESS Dataset
Licencia ISC	Permite usar, copiar, modificar y distribuir el software siempre y cuando se utilice el aviso de copyright y el aviso de permiso correspondiente	Librosa
Licencia BSD modificada	Permite usar y distribuir el software siempre y cuando se utilice el aviso de copyright y el aviso de permiso correspondiente y no se utilice los nombres de los creadores para promocionar un producto derivado sin el permiso de estos	Numpy
Licencia MIT	Permite usar, copiar, modificar, fusionar, publicar, sublicenciar, vender y distribuir el software siempre y cuando se utilice el aviso de copyright y el aviso de permiso correspondiente	Keras
Apache 2.0 open source license	Permite utilizar el software para cualquier propósito siempre y cuando se avise que se ha utilizado una licencia Apache	Tensorflow

TABLA 1.2. LICENCIAS UTILIZADAS EN EL PROYECTO (PARTE 2)

Creative Commons License Attribution-NonCommercial-ShareAlike	Permite combinar, adaptar y utilizar en el propio trabajo siempre y cuando se cite a los autores	Multiprocessing

1.4. Estructura del documento

Este documento contiene un total de 8 capítulos y un anexo, cuyo contenido se detalla a continuación:

- **Capítulo 1:** breve introducción donde se detallan el contexto y motivación del proyecto, su objetivo principal y los objetivos específicos, el marco regulador en el que se desarrolla, la estructura del documento y un listado de los acrónimos utilizados
- **Capítulo 2:** estado del arte donde se repasan los conocimientos pertinentes para el proyecto y trabajos similares que puedan servir como una base inicial
- **Capítulo 3:** definición del diseño que sigue el sistema desarrollado y el análisis realizado mediante requisitos y casos de uso
- **Capítulo 4:** definición del proceso de obtención y procesado de datos donde se especifican las fuentes de donde se extraen los conjuntos de datos, cómo se extraen y las técnicas utilizadas para procesarlos
- **Capítulo 5:** especificación del modelo de clasificación de emociones, del cual se detallan las arquitecturas, hiperparámetros y resultados de su experimentación; la elección del modelo final y la comparación entre distintas aproximaciones en cuanto al tratado de datos
- **Capítulo 6:** evaluación del sistema donde se detalla un plan de pruebas y los resultados obtenidos de aplicarlo
- **Capítulo 7:** establecimiento de la gestión del proyecto compuesta por la planificación, el presupuesto y el contexto socio-económico
- **Capítulo 8:** conclusiones donde se detallan los problemas encontrados y los posibles trabajos futuros
- **Anexo A:** resumen del proyecto completo en inglés

1.5. Acrónimos

- **ADAS:** Advanced Driving Assistance System, en castellano, Sistema Avanzado de Asistencia a la Conducción

- **API:** Application Programming Interface, en castellano, Interfaz de programación de aplicaciones
- **CV:** Cross Validation, en castellano, Validación Cruzada
- **DGT:** Dirección General de Tráfico
- **GPU:** Graphics Processing Unit, en castellano, Unidad de Procesamiento Gráfico
- **IA:** Inteligencia Artificial
- **IoT:** Internet of Things, en castellano, Internet de las Cosas
- **IVA:** Impuesto al Valor Añadido
- **MFCC:** Mel Frequency Cepstral Coefficients, en castellano, Coeficientes Cepstrales en las Frecuencias de Mel
- **PIB:** Producto Interior Bruto
- **RAVDESS:** The Ryerson Audio-Visual Database of Emotional Speech and Song
- **ReLU:** Rectified Linear Activation, en castellano, Unidad Lineal Rectificada
- **RNA:** Red de Neuronas Artificiales
- **SGD:** Stochastic Gradient Descent, en castellano, Gradiente Descendiente Estocástico
- **TPU:** Tensors Processing Unit, en castellano, Unidad de Procesamiento Tensorial

2. ESTADO DEL ARTE

2.1. Emociones

El concepto de emoción a lo largo de la historia ha estado basado en muchas teorías y no se ha alcanzado un consenso general en torno a su definición. Todas coinciden en que las emociones son estados que predisponen la manera en la que una persona se adapta al medio y a las situaciones en las que se encuentra, provocando una serie de manifestaciones apreciables denominadas sentimientos.

Las discrepancias entre investigadores se dan en torno a los factores que influyen en la definición de este este estado. Algunos como el profesor William James [9], defienden que solamente influyen los factores biológicos, mientras que otros como Walter Cannon [10] sostienen que el factor determinante es la cognición, al demostrar que mismos estados fisiológicos provocan emociones distintas. Al mismo tiempo, hay también una teoría planteada por Stanley Schacter que propone una combinación de las dos anteriores [11].

Otro foco de discusión con respecto a las emociones es la manera de clasificarlas, predominando dos teorías principales:

Según el psicólogo Paul Eckman [12], existen una serie de emociones básicas e interculturales que se clasifican de manera discreta en: ira, disgusto, miedo, felicidad, tristeza y sorpresa; categorizándolas así en una sola dimensión. Más adelante, su teoría evolucionó añadiendo una serie de emociones secundarias resultantes de la combinación de las emociones básicas y demostrando una fuerte relación entre ellas y las expresiones faciales como se observa en la Fig. 2.1.

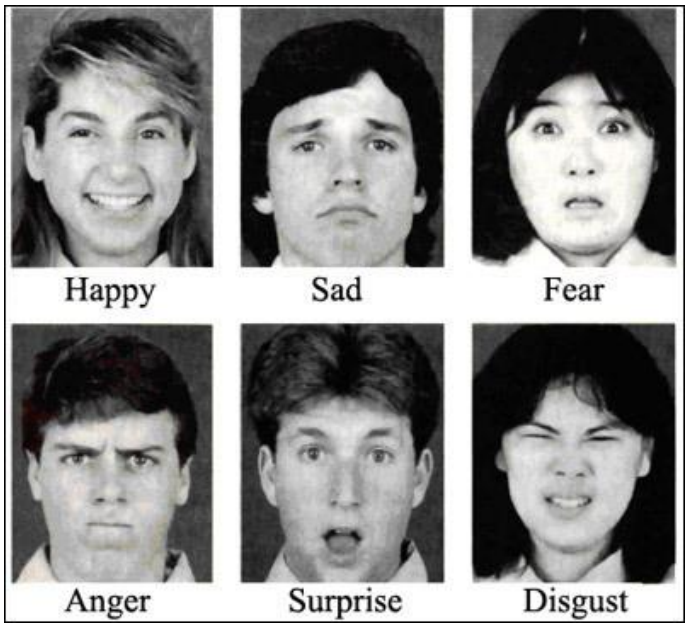


Fig. 2.1. Expresiones faciales asociadas a cada emoción [12]

Una aproximación distinta es la sugerida por el también psicólogo James A. Russell [13], que presenta un modelo circumplejo en el cual las emociones se clasifican con respecto a dos dimensiones: valencia y excitación. El valor de la valencia especifica el grado de negatividad o positividad de la emoción mientras que el valor de excitación hace referencia a la intensidad de la misma. Dentro de este espacio bidimensional se van encontrando diferentes emociones de cualquier tipo como se muestra en la Fig. 2.2.

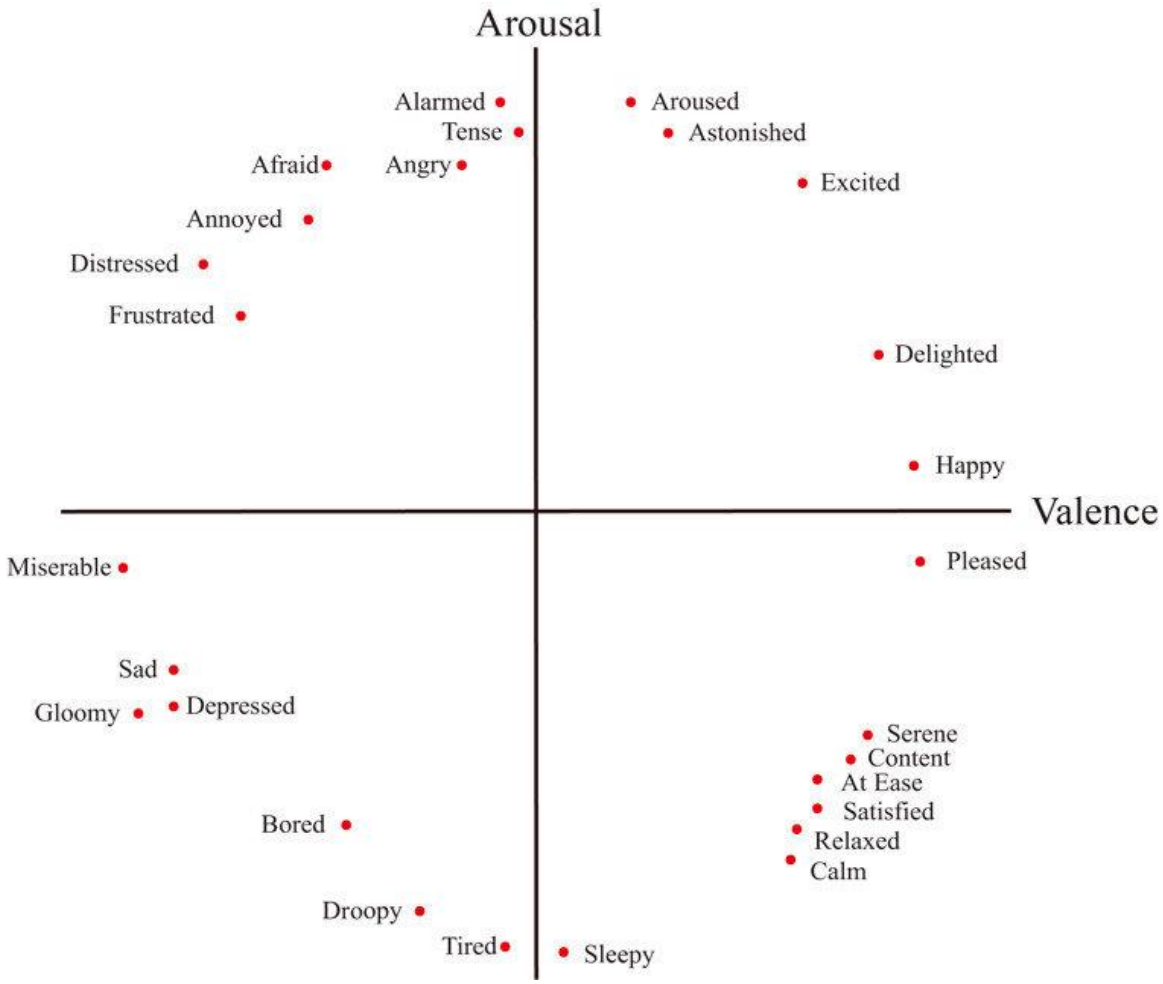


Fig. 2.2. Modelo circumplejo de clasificación de emociones según James A. Russell [13]

2.1.1. Emociones en el habla

El ser humano tiene la capacidad de comunicarse oralmente mediante la palabra. En este proceso de la comunicación intervienen una gran cantidad de factores como pueden ser el emisor, receptor, mensaje, código o contexto; los cuales a su vez son influenciados por la emoción de las personas involucradas en la comunicación.

La afirmación de que las emociones afectan al proceso de la comunicación se puede obtener fácilmente del hecho de que la comunicación oral depende de factores fisiológicos, como pueden ser el estado de los órganos del aparato fonador, los cuales dependen de la emoción del sujeto en cuestión [14].

Estas emociones, que se manifiestan como sentimientos en el habla, tienen un papel fundamental en la comunicación, ya que influyen en la interpretación que hacen los receptores del significado del mensaje transmitido. Como por ejemplo en la frase: “Me tienes de buen humor”, la cual asociada a una emoción de felicidad transmite un mensaje positivo mientras que asociada a una emoción de enfado transmite un mensaje de reprimenda.

Los estudios tanto pasados como actuales han centrado sus esfuerzos en identificar cuáles son esos factores universales que asocian el habla con una determinada emoción, para poder así inducirla mediante un procedimiento específico. Esta tarea presenta un gran desafío, ya que pretende buscar factores diferenciables para cualquier caso de un elemento con innumerables características diferentes y que, además, va evolucionando con el paso del tiempo por motivos biológicos y culturales.

Un estudio de relevancia en esta tarea fue el realizado por los doctores en psicología Klaus R. Scherer y Tom Johnstone, los cuales mediante la observación de tres tipos distintos de experimentos según como se provocaba la emoción en el habla: de manera natural, inducida mediante fármacos psicoactivos o actuada; obtuvieron una serie de relaciones entre ciertas características del habla y su relación con una serie de emociones básicas [15], como se puede observar en la Fig. 2.3, la cual muestra una matriz donde las columnas son las distintas emociones y las filas son ciertas características relacionadas con el sonido como las intensidad o la frecuencia. Los valores de las casillas toman el símbolo de una flecha ascendente en aquellas en las que un aumento de la característica del sonido va asociado a la emoción; y una flecha descendente cuando va relacionado con un descenso en la valencia de esa característica.

	Stress	Anger/rage	Fear/panic	Sadness	Joy/elation	Boredom
Intensity	↗	↗	↗	↘	↗	
F0 floor/mean	↗	↗	↗	↘	↗	
F0 variability		↗		↘	↗	↘
F0 range		↗	↗(↘)	↘	↗	↘
Sentence contours		↘		↘		
High frequency energy		↗	↗	↘	(↗)	
Speech and articulation rate		↗	↗	↘	(↗)	↘

Fig. 2.3. Matriz de relación entre características del sonido y ciertas emociones [15]

A pesar de las críticas por una aproximación reduccionista en cuanto a las distintas emociones y características que se tuvieron en cuenta en el estudio, este sirve como base de la que partir a la hora de afirmar que existen características identificables en el habla que permiten asociar una cierta emoción.

2.1.2. Influencia de las emociones en la conducción

Los estados emocionales afectan directamente a la manera en la que los seres humanos actúan y se desenvuelven en el día a día. Esto incluye también las actividades que implican conducir un automóvil.

Un factor fundamental en la conducción es la atención, ya que una distracción a velocidades elevadas provoca que la distancia de reacción sea en muchas ocasiones incontrolable, y acabe provocando una gran cantidad de accidentes de tráfico. Aunque la atención no se considera una emoción, esta se ve positiva o negativamente influida por las distintas emociones del individuo [16]. Por lo tanto, se puede concluir que las emociones afectan en la conducción de diversas maneras.

Un estudio realizado en la Universidad Católica de Eichstätt-Ingolstadt por K. Steinhauser et al. [17] investigó los efectos de inducir distintas emociones, tanto positivas como negativas, en conductores que realizaban distintas tareas relacionadas con la conducción. Entre los resultados obtenidos relevantes se encuentran:

- La inducción de las emociones mediante el uso de la imaginación o la música tiene efectos directos sobre el estado anímico del conductor y sobre su conducción

- Hay algunas emociones que aumentan el efecto sobre el tiempo de reacción del conductor cuando éste está realizando tareas que requieren de su atención, como por ejemplo el aumento del tiempo de reacción en el uso de los frenos para emociones como calma o alegría y su reducción en el uso del acelerador

- Ciertas emociones tienen influencia sobre determinadas tareas de la conducción mientras que apenas muestran cambios sobre otras. Por ejemplo, en la Fig. 2.4 se muestra como en los experimentos realizados, la ira tiene un gran efecto sobre el tiempo de reacción en la tarea de acelerar pero el cambio sobre la tarea de frenar es mínimo.

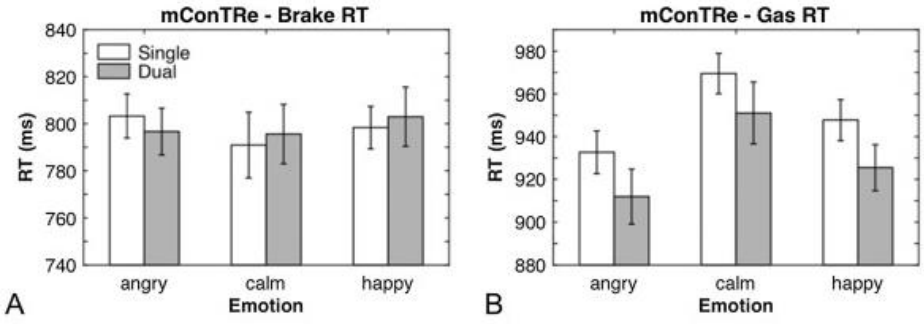


Fig. 2.4. Gráficas de la relación entre los tiempos de reacción en la utilización de frenos y acelerador para ciertas emociones [17]

2.2. Sistemas de ayuda a la conducción

Los sistemas avanzados de ayuda a la conducción (ADAS) surgen de la necesidad de ayudar al conductor en ciertos aspectos de la conducción y a su vez, reducir o mitigar los accidentes de tráfico provocados por fallos humanos. Mediante el uso de una variedad de sensores y ordenadores de abordo que procesan la información y toman decisiones, se consigue introducir un cierto nivel de automatización que asiste y protege a los usuarios de un automóvil.

Este grado de automatización no es el mismo para todos los ADAS y a partir de cierto punto, se dejan de considerar sistemas de asistencia y se consideran vehículos autónomos. Para poder diferenciarlos, se explican a continuación los distintos niveles según la organización de estándares SAE International [18] y que se pueden encontrar en la Fig. 2.5.

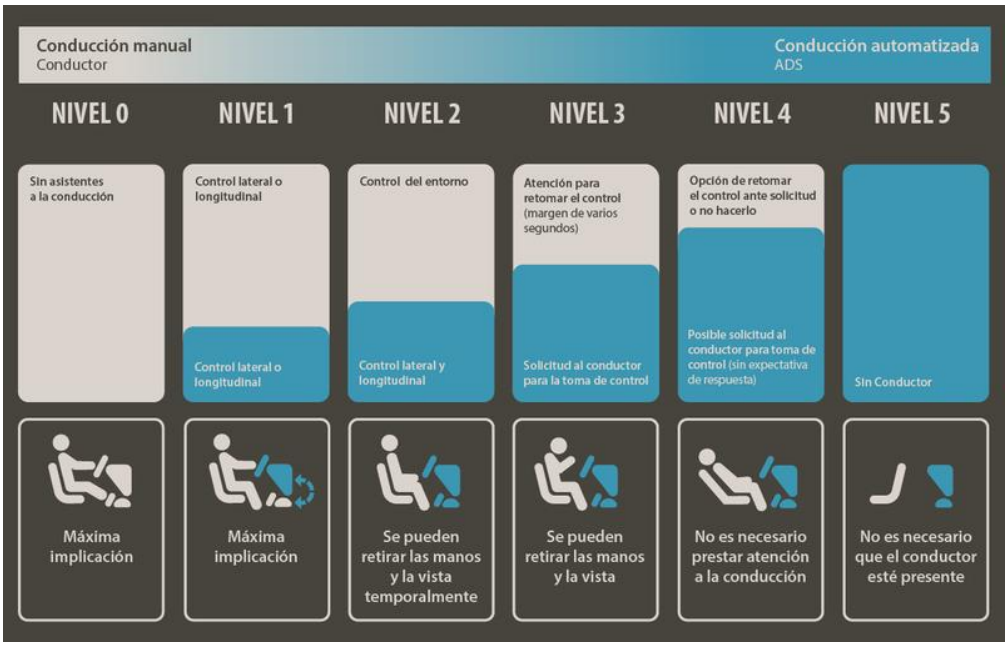


Fig. 2.5. Seis niveles de automatización de un vehículo [19]

- **Nivel 0:** No existe ningún tipo de automatización. Todas las actividades relacionadas con la conducción son realizadas por el conductor.

- **Nivel 1:** Asistencia al conductor. Se automatiza el movimiento longitudinal o transversal del vehículo, pero no ambos a la vez, y se espera que el conductor realice el resto de las actividades relacionadas con la conducción. En este grupo se encuentran sistemas como el mantenimiento de carril, frenado de emergencia o asistencia al aparcamiento.

- **Nivel 2:** Automatización parcial de la conducción. Se automatiza de manera continuada tanto el movimiento longitudinal como transversal del vehículo y requiere que el conductor realice el resto de tareas y supervise su funcionamiento. En este grupo se encuentran sistemas como el cambio de carril o aparcamiento.

A partir de este nivel entran los denominados comúnmente como vehículos autónomos:

- **Nivel 3:** Automatización de la conducción condicional. Se automatiza una tarea específica de la conducción, pero el sistema puede requerir que el conductor tome el control en un determinado momento.

- **Nivel 4:** Automatización de la conducción alta. Se automatiza una tarea específica de la conducción y el sistema no requiere en ningún momento la intervención del conductor.

- **Nivel 5:** Automatización completa. Se automatizan todas las tareas pertinentes a la conducción del vehículo. Esto significa que el vehículo es capaz de circular sin que el conductor esté en él.

Toda esta automatización no sería posible sin una serie de sensores los cuales se combinan en los distintos vehículos y que muchas veces forman parte de los ADAS. A continuación se definen los más comunes [20]:

- **Radar:** tecnología que mide las distancias a los objetos dependiendo de lo que tardan en volver las ondas de radio emitidas. Destaca por su robustez, precio asequible y capacidad de funcionar en condiciones atmosféricas desfavorables. Su principal desventaja es su incapacidad de reconocer qué objeto es el que está detectando.

- **Ultrasonidos:** tecnología que mide las distancias a los objetos dependiendo de lo que tardan en volver las ondas de ultrasonido emitidas. Destaca por su fiabilidad, robustez y precio. Su principal desventaja es su corto alcance.

- **Lidar:** tecnología que mide las distancias mediante un láser y a partir del cual crea una imagen en tres dimensiones del escenario. Destaca por su efectividad y su gran alcance de hasta 300 metros. Su principal desventaja es su precio inasequible para su utilización en coches comercializados.

- **Cámara:** tecnología que mediante el uso de una o más cámaras es capaz de monitorizar lo que ocurre en el entorno del vehículo. Destacan por su bajo coste y su capacidad de distinguir formas y colores. Su principal desventaja es su incapacidad de funcionar en condiciones atmosféricas que limiten la visibilidad.

La combinación de estos sensores y muchos otros, combinados con técnicas de automatización y computación son lo que constituyen la mayoría de los ADAS. Para poder diferenciarlos según su función, la empresa tecnológica de soluciones Internet of Things (IoT) Samsara presenta los siguientes grupos [21]:

- **Adaptativos:** realizan pequeños ajustes en el vehículo de forma autónoma para mejorar la seguridad de la conducción a partir de los datos del entorno.

- **Automatizados:** toman el control del vehículo de forma autónoma para evitar un accidente inminente.
- **De monitorización:** visualizan el entorno para ofrecer información detallada de manera autónoma sobre las condiciones del mismo.
- **Alertadores:** sistemas automatizados que alertan al conductor en tiempo real de un posible peligro en la conducción.

2.3. Sonido

La audición es uno de los cinco sentidos que tienen los seres humanos y desempeña un papel fundamental en la comprensión y percepción del entorno. Además, es uno de los pilares fundamentales de la comunicación oral.

Para poder estudiar y analizar el sonido es importante comprender los principios físicos que lo definen. El sonido es una serie de ondas que se transmiten por un medio elástico que, en el caso de la comunicación oral es el aire. Estas ondas están definidas por una serie de características las cuales se muestran en la Fig. 2.6.

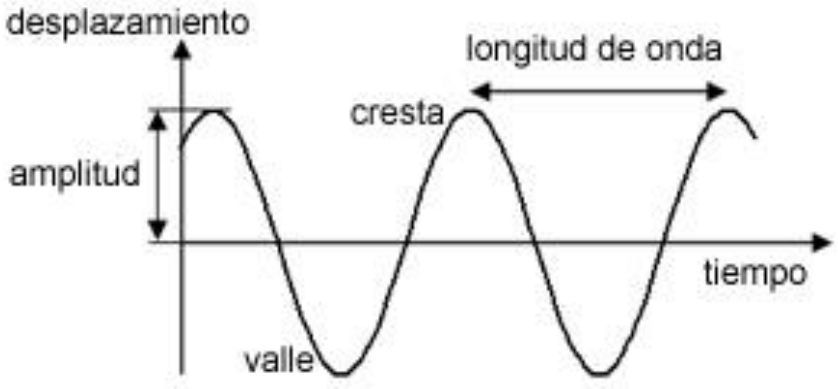


Fig. 2.6. Partes de una onda [22]

Las definiciones de estas partes son las siguientes:

- **Ejes:** se cuenta con dos ejes. El eje vertical representa el desplazamiento de la onda mientras que el horizontal representa el tiempo transcurrido.
- **Cresta y Valle:** punto de la onda más alejado del eje horizontal, positivo en el caso de la cresta y negativo en el caso del valle.
- **Amplitud:** distancia de una cresta o un valle al eje horizontal. Este valor determina el volumen del sonido.
- **Longitud de onda:** distancia entre las crestas o valles de dos ciclos consecutivos.
- **Periodo:** tiempo que tarda en completarse un ciclo completo.

- **Frecuencia:** número de ciclos que realiza la onda en un segundo. Es la inversa del periodo.

A parte de estas cualidades, a las ondas sonoras también tienen una serie de características asociadas a la percepción humana de las mismas [23]:

- **Intensidad:** depende de la presión, frecuencia y duración de la onda. Permite diferenciar entre sonidos fuertes y débiles.

- **Tono:** depende principalmente de la frecuencia y permite diferenciar entre un sonido agudo y uno grave.

- **Timbre:** depende principalmente de las ondas que se combinan para formar ese sonido. Permite diferenciar dos sonidos con el mismo tono e intensidad pero que provienen de focos distintos.

Una vez definida una onda, es pertinente saber que el sonido rara vez está formado por una sola onda, si no que se trata de la combinación de muchas de ellas, las cuales combinan sus propiedades para formar una nueva. Más adelante en este documento se presentarán formas de extraer las ondas puras (que solamente están formadas por una) para así poder analizarlas individualmente.

2.3.1. Digitalización del sonido

En el contexto de utilizar el sonido en sistemas informáticos como por ejemplo el caso de un ADAS, el sonido necesita ser transformado de tal forma que sea procesable por un ordenador, es decir, en unos y ceros.

Los micrófonos son los dispositivos encargados de transformar la presión del aire provocada por las ondas en una señal analógica. El problema de esta transformación surge que la naturaleza continua de esa señal no es tratable por un ordenador. Es por eso por lo que se utiliza un método llamado muestreo o sampling.

El muestreo o *sampling* consiste en obtener la amplitud de la onda en un instante específico de manera periódica. El número de muestras que se obtienen por segundo viene determinado por la frecuencia de muestreo o *sampling rate*. A mayor frecuencia de muestreo se obtendrán pistas de audio de mayor calidad pero que necesitarán de mayores recursos de almacenamiento.

Tras realizar el muestreo es necesario transformar ese valor de voltaje obtenido por el micrófono a una serie de números reales que representen el valor la amplitud de la onda. Esto se consigue mediante un proceso denominado cuantificación, el cual consiste en truncar o aproximar los valores para poder almacenarlos en tamaños de 8 a 16 bits normalmente.

Una vez terminada esa cuantificación, se procede a realizar el último paso denominado codificación, y que básicamente consiste en transformar los valores obtenidos en la cuantificación a valores en binario representado por unos y ceros.

Todo este proceso se muestra en la Fig. 2.7.

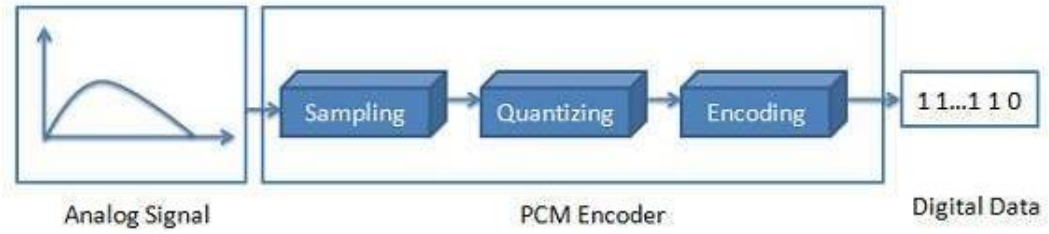


Fig. 2.7. Proceso de digitalización del sonido [24]

2.4. Inteligencia Artificial

La empresa tecnológica IBM aporta la siguiente definición de Inteligencia Artificial (IA): *"Artificial intelligence leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind."* [25]. Que se traduce como: "La Inteligencia Artificial aprovecha los ordenadores y las máquinas para imitar la resolución de problemas y la capacidad de toma de decisiones de la mente humana."

Esta definición aportada por IBM marca una clara relación entre la inteligencia artificial y la inteligencia humana a la cual intenta replicar. Esta relación no es innovadora ya que fue el profesor de la universidad de Stanford, John McCarthy, quien estableció que no existía una definición sólida de Inteligencia Artificial que no hiciera referencia a la inteligencia humana [26].

Debido a la estrecha relación entre la Inteligencia Artificial y la inteligencia humana es necesario establecer una definición de inteligencia humana. El diccionario de Oxford lo describe como la "*facultad de la mente que permite aprender, entender, razonar, tomar decisiones y formarse una idea determinada de la realidad*" [27].

Una manera de crear una división en los distintos paradigmas de Inteligencia Artificial que se pueden utilizar para resolver un problema, es la aportada por F. Corea [28] el cual propone las siguientes divisiones:

- **Basadas en la lógica:** representan conocimiento y procedimientos.
- **Basadas en el conocimiento:** basadas en grandes bases de datos, reglas e inferencias.
- **Métodos probabilísticos:** para la toma de decisiones en entornos incompletos.
- **Aprendizaje automático:** aprenden al exponerles los datos de las instancias de un problema.

- ***Embodied interaction*:** involucra características del cuerpo humano en la definición de inteligencia.
- **Búsqueda y optimización:** para realizar búsquedas de una manera inteligente y eficiente.

En conclusión, la Inteligencia Artificial es una disciplina con infinidad de aplicaciones muy diversas entre sí capaces de resolver gran cantidad de problemas ya que su homóloga humana tiene esa misma versatilidad y potencia.

2.5. Aprendizaje Automático

De las distintas aproximaciones para abordar un problema que ofrece la Inteligencia Artificial, cabe destacar el Aprendizaje Automático debido a su aplicación en este proyecto. La relación entre la IA y el Aprendizaje Automático, o en inglés *Machine Learning*, se muestra en la Fig. 2.8, donde además se incluye el subconjunto de ambos denominado *Deep Learning* o Aprendizaje Profundo en castellano.

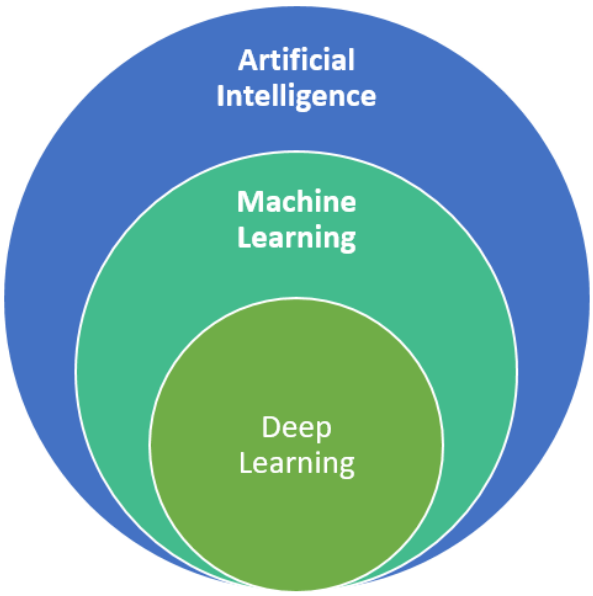


Fig. 2.8. Relación entre Inteligencia Artificial, Aprendizaje Automático y Aprendizaje Profundo [29]

El Aprendizaje Automático es, por lo tanto, un subconjunto de la IA el cual se caracteriza por la capacidad que tienen los algoritmos de aprender a realizar predicciones, la cual obtienen a través de la exposición de un conjunto de datos de un determinado problema que el algoritmo aprende a generalizar. Lo característico de este tipo de algoritmos y la razón de su popularidad es que no es necesario que el diseñador especifique un conjunto de pasos específicos que se tienen que seguir para solucionar el problema, si no que es el propio algoritmo el que aprende a realizar ese proceso necesario para llegar a una solución.

2.6. Redes de Neuronas Artificiales

Las Redes de Neuronas Artificiales (RNA) es una técnica de IA que se encuentra dentro del subgrupo de técnicas de Aprendizaje Automático, y que basa su arquitectura y funcionamiento en las neuronas biológicas humanas.

En la Fig. 2.9 se pueden observar las distintas partes de una neurona humana.

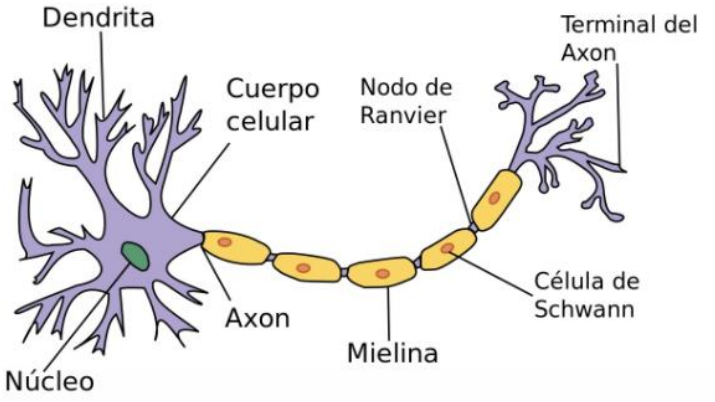


Fig. 2.9. Partes de la neurona humana [30]

Las partes que tienen un equivalente en la arquitectura de una neurona artificial son las siguientes [31]:

- **Dendritas:** su función es recibir los impulsos eléctricos que llegan de otras neuronas y transmitirlos al núcleo.
- **Núcleo:** se encarga de combinar todos los impulsos obtenidos de las dendritas y transmitirlo por el axón en caso de que se active.
- **Axón:** se encarga de transportar los impulsos eléctricos hasta sus terminaciones que se transmitirán a través de la sinapsis a otras neuronas.
- **Sinapsis:** reacción a través de un neurotransmisor que comunica el impulso eléctrico de una neurona a otra colindante.

Estas partes se pueden encontrar en la arquitectura diseñada por Frank Rosenblatt, el cual se basó en la primera neurona artificial diseñada por Warren McCulloch y Walter Pitts en 1943 para diseñar el Perceptrón, que se muestra en la Fig. 2.10.

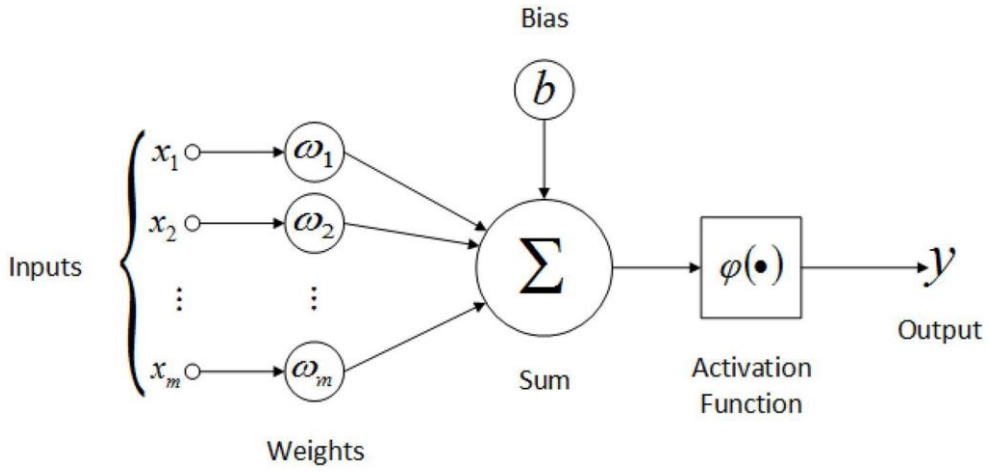


Fig. 2.10. Partes de una neurona artificial [32]

- **Entradas:** las entradas por donde se recogen los valores $X_1, X_2, X_3, \dots, X_m$ equivalen a las dendritas en la neurona humana.
- **Salida:** la salida de la neurona por donde se devuelve el valor Y_k se corresponde con el axón en la neurona humana.
- **Combinación de las entradas y función de activación:** las funciones que combinan todos los valores de entrada mediante una combinación lineal y su posterior activación dependiendo de una función $\varphi(\cdot)$ se corresponden con el núcleo de la neurona humana. Este proceso se muestra en la siguiente fórmula:

$$Y_k = \varphi\left(b + \sum_{i=1}^{m} w_i \cdot X_i\right), donde\ \varphi(x) = \begin{cases} 1, & si\ x \geq 0 \\ -1, & si\ x < 0 \end{cases}$$

  Donde φ es la función de activación, b es el *bias*, w_i los pesos de la neurona artificial y X_i los valores de entrada.

Esta definición de neurona es capaz de realizar una clasificación binaria dependiendo de los atributos de los datos de entrada. Sin embargo, no es suficiente para considerarlo un algoritmo dentro del subconjunto del Aprendizaje Automático, ya que es necesario que sea el propio algoritmo el que realice el ajuste de pesos propios del aprendizaje. Es por eso por lo que también se define la regla de aprendizaje del Perceptrón, la cual varía los pesos de las conexiones de la entrada al cuerpo de la neurona si la clasificación es errónea utilizando la siguiente fórmula:

$$\Delta\, w_i = d(x) \cdot x_i\,,\ \ donde\ d(x)\ es\ la\ clase\ de\ la\ instancia\ clasificada$$

Esta fórmula de aprendizaje se modifica en el modelo ADALINE creado por Bernard Widrow y Ted Hoff en 1960, el cual mediante la aplicación de la regla Delta permitía ajustar el grado de modificación de pesos dependiendo de la desviación de la predicción.

Tanto el Perceptrón como el ADALINE son capaces de realizar clasificaciones de conjuntos de datos linealmente separables, es decir que pueden ser separados por una

recta, pero no eran funcionales para conjuntos no linealmente separables como el que se muestra en la Fig. 2.11.

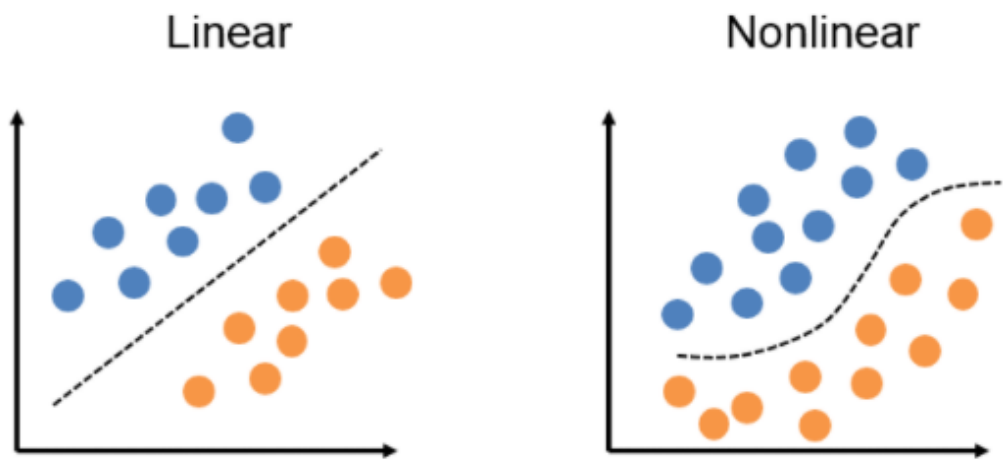


Fig. 2.11. Diferencias entre conjuntos linealmente y no linealmente separables [33]

Esta falta de un algoritmo capaz de generalizar problemas no linealmente separables se soluciona en 1975 con la aparición del Perceptrón Multicapa, el cual mediante la introducción de más de una capa de neuronas y la utilización de funciones de activación no lineales resuelve el reto de los conjuntos no linealmente separables. Sin embargo, el proceso de aprendizaje de esta arquitectura se tenía que realizar a mano por los diseñadores hasta la aparición de la regla delta generalizada a manos de Rumelhart, Hinton y Williams [34], también conocido como *Backtracking*, la cual permitía realizar la técnica del descenso de gradiente a redes de neuronas con múltiples capas de neuronas.

Estas arquitecturas multicapa tienen una serie de elementos comunes:

- **Capa de entrada:** primera capa de neuronas a la cual son introducidos los datos.
- **Capas ocultas:** serie de capas conectadas en paralelo.
- **Capa de salida:** última capa que produce el output de la red.

Un ejemplo de red multicapa se puede observar en la Fig. 2.12 donde se presenta una red densamente conectada con dos capas ocultas.

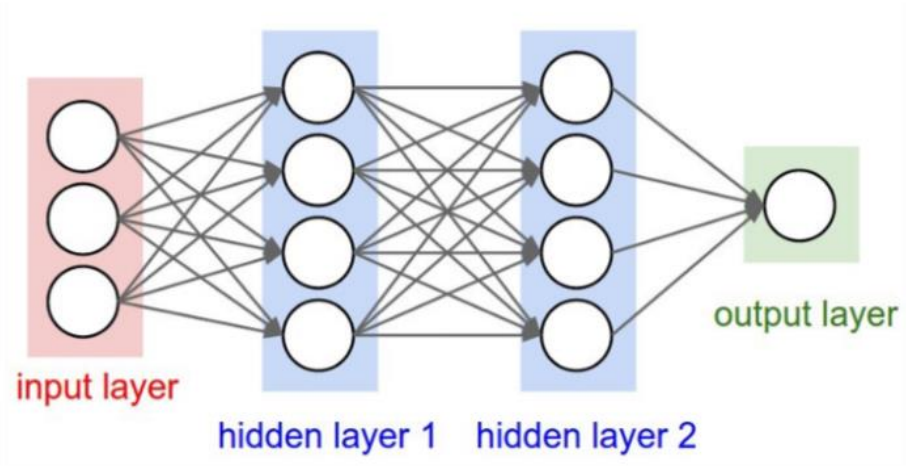


Fig. 2.12. Ejemplo de red de neuronas multicapa [32]

2.6.1. Deep Learning: Redes Convolucionales

Con la aparición del algoritmo de *Backtracking*, se presentó la oportunidad de crear modelos con gran cantidad de neuronas y capas capaces de aprender tareas complejas, sin embargo, estos resultaron muy complejos de entrenar.

No fue hasta 2006, cuando tras la aparición y desarrollos de chips optimizados para las operaciones de matrices, GPUs y TPUs, se fue capaz de entrenar esos modelos con un número masivo de neuronas y capas, las cuales entran dentro del subconjunto denominado Aprendizaje Profundo o *Deep Learning.*

En concreto cabe destacar las Redes Convolucionales debido a su utilización en este proyecto. Esta arquitectura está formada por dos elementos principales: las capas convolucionales y un perceptrón multicapa densamente conectado.

Las capas convolucionales son las encargadas de extraer las características específicas de una matriz que más tarde el perceptrón multicapa utiliza para clasificar las instancias. Tomando como ejemplo una imagen en RGB como se muestra en la Fig. 2.13, y que se representa como tres matrices de dos dimensiones, se explican las dos transformaciones asociadas a las capas convolucionales: convolución y *pooling.*

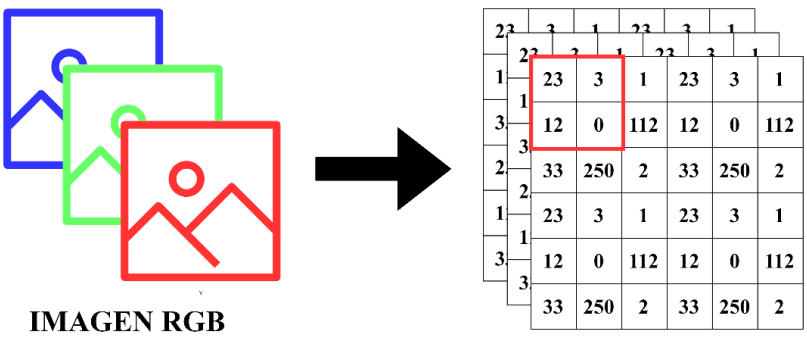


Fig. 2.13. Representación matricial numérica de una imagen RGB de tres canales

La transformación de convolución consiste en ir desplazando una matriz de valores llamada *kernel* por la imagen tanto verticalmente como horizontalmente, e ir calculando el valor de la combinación lineal de estos valores tal y como se muestra en el ejemplo de la Fig. 2.14, en la cual se aplica a la ventana marcada en rojo de la Fig. 2.13 un *kernel* de tamaño 2x2 para obtener un único valor.

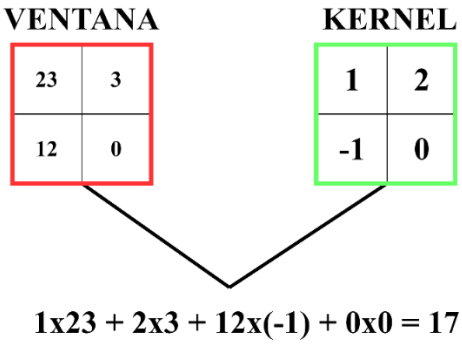


Fig. 2.14. Aplicación del *kernel* a una ventana de la imagen

Al aplicar el *kernel* a toda la imagen se obtiene una nueva matriz de valores a la cual se le puede aplicar la segunda transformación denominada *pooling*, cuya función es reducir el tamaño de la imagen seleccionando solamente los valores más importantes. Por ejemplo, en la Fig. 2.15 se muestra el proceso del *Max Pooling* de tamaño 2x2 en la cual se obtiene el valor máximo para cada ventana de distinto color, aunque también existen otros tipos de *pooling* que seleccionan el valor mínimo o el valor medio entre muchos otros.

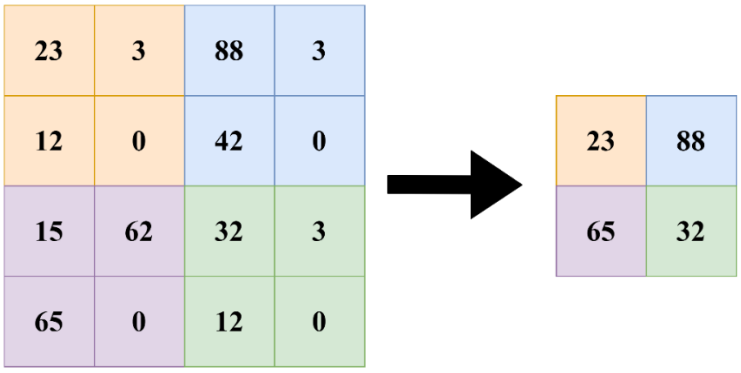


Fig. 2.15. Aplicación de filtro *Max Pooling*

Estas dos transformaciones se van aplicando secuencialmente hasta llegar al perceptrón multicapa densamente conectado al final de la red. Este se encarga de analizar todas las características extraídas por las capas de convolución y realizar la clasificación o regresión necesaria.

2.7. Trabajos similares

Para poder tener una visión más actual y práctica de la situación, se estudian las aproximaciones y métodos utilizados por otros estudios relacionados. Esta revisión de

trabajos similares es especialmente útil al trabajar con Redes Convolucionales ya que, al tener este tipo de arquitecturas una gran cantidad de parámetros ajustables se puede ahorrar tiempo de experimentación, al descartar o seguir los experimentos realizados por un estudio similar.

2.7.1. Reconocimiento de la emoción a través del habla mediante *Machine Learning*

Este trabajo realizado por M. Wadhwa, A. Gupta y P. K. Pandey [35] en el Instituto Indio de Tecnología, presenta una solución basada en Redes Convolucionales para clasificar por emoción según el sentimiento que transmitían pistas de audio que con personas hablando.

Al igual que la mayoría de los proyectos de minería de datos este divide su estructura en: análisis del problema, obtención de los datos, preprocesado de los datos, creación de los modelos y evaluación.

El análisis del problema aporta información relevante como la dificultad de la clasificación de emociones debido a sus grandes diferencias de una persona a otra o aspectos humanos como puede ser el sarcasmo. También presenta la posibilidad de clasificar emociones de dos maneras distintas: una clasificación discreta, la cual clasifica las pistas con etiquetas que representan cada una de las emociones, o una clasificación dimensional donde se clasifica una pista mediante el grado de pertenencia de todas las emociones.

La obtención de datos se realiza a partir de cinco fuentes de datos distintas. Al unificarlas todas se consigue una base de datos con representación de distintos idiomas, edades, géneros, duraciones y frecuencias de muestreo. Es importante destacar también que las emociones se obtienen de actores profesionales a los que se les graba recreando las emociones.

En el preprocesado de datos se parte de archivos *.wav* que no tienen ningún tipo de compresión y contienen el valor de la amplitud para valores de tiempo especificados por la frecuencia de muestreo. Dado que las distintas fuentes de datos tienen distintas frecuencias de muestreo, se aplican transformaciones para unificar todo el dataset a un valor. Una vez realizada esta transformación, se utilizan distintos filtros para obtener representaciones de los datos más ricas en información para que puedan ser posteriormente presentadas a la red convolucional. Estas transformaciones son la obtención de medidas referentes al sonido como la energía o la obtención de un espectrograma de frecuencias que aporta más información que la pista original. En último lugar, se realiza una selección de características para descartar aquellas que no aportan información relevante de cara a la clasificación.

Tras procesar los datos comienzan la fase de experimentación donde prueban dos aproximaciones distintas de redes convolucionales, una con una entrada de una dimensión

y otra de dos dimensiones, y otras técnicas de *Machine Learning* como *Support Vector Machines* o *XGBoost.*

Finalmente, se evalúa el modelo utilizando la métrica de la precisión y se obtiene un 75% de acierto en el conjunto de test para el mejor modelo, que es el de redes convolucionales con entradas de una dimensión.

2.7.2. Análisis del rendimiento de reconocimiento de emociones acústico para interfaces conversacionales en automóviles

El trabajo realizado por C. M. Jones y I-M. Jonsson [36] surge de la proliferación de asistentes con una interfaz conversacional en el ámbito de la conducción, y de la posibilidad de mejorar su desempeño con la introducción de análisis de sentimiento para que se puedan adaptar al estado anímico del conductor.

La base de datos de este proyecto fue generada mediante sesiones de conducción donde sujetos de prueba realizaban una serie de tareas y luego mediante un cuestionario reportaban que emoción tenían en cada instante.

Tras el procesado de datos y la evaluación de muestras con una duración de dos segundos, se obtuvo una tasa de acierto del 65%, siendo el 15 % de las que se fallaron emociones equivalentes para el contexto de la conducción.

La conclusión en este trabajo es confirmar la viabilidad de crear sistemas de asistencia por voz que comprendan las emociones del usuario y actúen en consecuencia para mejorar su experiencia; y que en caso de que se equivoquen en la predicción, no se verá afectada negativamente la seguridad del conductor y los pasajeros.

2.8. Aportaciones al estado del arte

Este proyecto se desarrolla dentro un contexto que se ha definido en este apartado del estado del arte. Esta puesta en contexto sirve para aportar información y posibles guías a seguir en el proceso de solucionar el problema planteado.

Este problema en concreto concierne a muchas partes que se han desarrollado previamente como pueden ser: las emociones humanas, los ADAS, el sonido o la Inteligencia Artificial.

Las aportaciones al estado del arte que realiza este proyecto son sobre todo orientadas a la metodología a la hora de tratar los datos que utilizan, ya que lo común no es trabajar con imágenes si no con vectores directamente.

A parte de esta metodología, también se aporta el diseñar la componente de un sistema específicamente orientado a la asistencia en carretera. Entre las cuales se podrían obtener aplicaciones para alertar al conductor o incluso adaptar el resto de sistemas de asistencia al estado anímico del usuario.

3. DISEÑO Y ANÁLISIS DEL SISTEMA

3.1. Arquitectura del sistema

El diseño del sistema en el que integrar la componente desarrollada consta de tres elementos principales: el micrófono situado en el interior del vehículo, el cual se encarga de obtener los datos respectivos al habla del conductor, el sistema de procesado, que se encarga de predecir la emoción del conductor y el cual se diseña en este proyecto, y el sistema actuador, que se encarga de tomar las medidas de seguridad oportunas. A pesar de que no se diseña todo el sistema, es necesario especificar su arquitectura para poder diseñar la componente acorde a este.

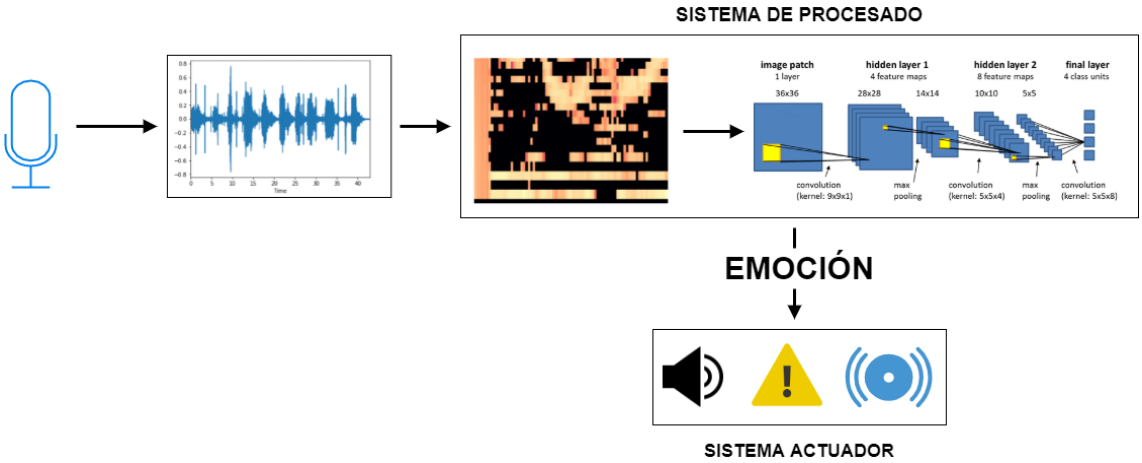


Fig. 3.1. Esquema de la arquitectura del sistema donde se integra la componente diseñada

En la Fig. 3.1 se muestra un diagrama de la arquitectura del sistema. En la que se encuentran las tres componentes principales conectadas de manera secuencial.

En primer lugar, el micrófono transforma las ondas de sonido que viajan por el medio en una señal digital que es capaz de procesar un ordenador. En concreto esta señal tiene un formato *.wav*, la cual tienen los datos de amplitud de cada muestra temporal específica y que no cuenta con ningún tipo de compresión para no perder información que puede ser relevante para el sistema de procesado.

Después de obtener el archivo de sonido, este accede al sistema de procesado en donde pasa por dos etapas distintas. La primera convierte el archivo *.wav* en distintas imágenes que representan su espectrograma de Mel del cual se hablará más adelante y a continuación es procesada por la red de neuronas convolucional, la cual produce una predicción sobre la emoción correspondiente al habla del conductor.

Una vez obtenida la emoción, el sistema actuador utiliza esa información para actuar en consecuencia. Entre esas actuaciones podrían encontrarse alertas visuales, auditivas o incluso la adaptación de otros sistemas de asistencias a esa emoción en concreto. No se especifica más acerca de este sistema ya que no entra dentro del alcance de este proyecto y requeriría la colaboración de especialistas en controlar las emociones de las personas mediante agentes externos.

3.2. Tecnologías utilizadas

Todo el desarrollo del sistema de procesado ha sido realizado en el lenguaje de programación Python debido a la existencia de librerías de utilidad y la existencia de experiencia previa en este lenguaje. Las librerías externas que se han utilizado son las siguientes:

- **Librosa:** permite trabajar con pistas de audio y música y ofrece una serie de funciones preestablecidas que permiten modificarlas y extraer información de ellas. En este proyecto se ha utilizado para procesar las pistas de audio, aplicarle filtros para eliminar partes en silencio y crear los espectrogramas que se transforman a imágenes.

- **Numpy:** permite realizar cálculos numéricos eficientes sobre todo en vectores y matrices de diferentes dimensiones. En este proyecto se ha utilizado para manipular en primera instancia los vectores de los archivos *.wav* en el proceso de extracción y etiquetado.

- **Multiprocessing:** esta librería ofrece una interfaz para la creación de distintos hilos de trabajo que permiten paralelizar procesos. Se ha utilizado para paralelizar la transformación de los archivos *.wav* en imágenes ya que requería de una gran capacidad de computación que no era posible alcanzar en una ejecución de un único hilo.

- **Keras y Tensorflow:** estas librerías ofrecen una interfaz sencilla y rápida para la experimentación y creación de modelos de redes de neuronas. En este proyecto se han utilizado para experimentar con distintos modelos capaces de generalizar los datos generados.

El desarrollo y entrenamiento de los modelos de redes de neuronas requieren de una gran cantidad de memoria y capacidad de procesamiento de matrices propias de las tarjetas gráficas (GPU). El acceso a estos dispositivos se ha realizado de manera remota por lo que el proceso de visualización de gráficas es bastante limitado. Por este motivo se ha utilizado una tecnología denominada CometML, la cual permite mediante una conexión a través de una Interfaz de Programación de Aplicaciones (API) visualizar una gran variedad de datos en tiempo real en el navegador web como se muestra en la Fig. 3.2, donde se listan los distintos experimentos realizados o en la Fig. 3.3 donde se muestran las gráficas y datos que se pueden consultar de cada experimento realizado.

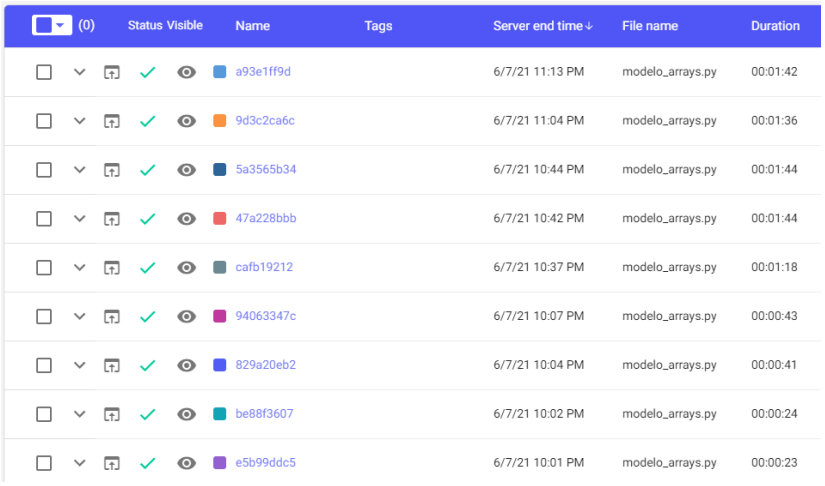


Fig. 3.2. Menú de experimentación de CometML con listado de experimentos realizados

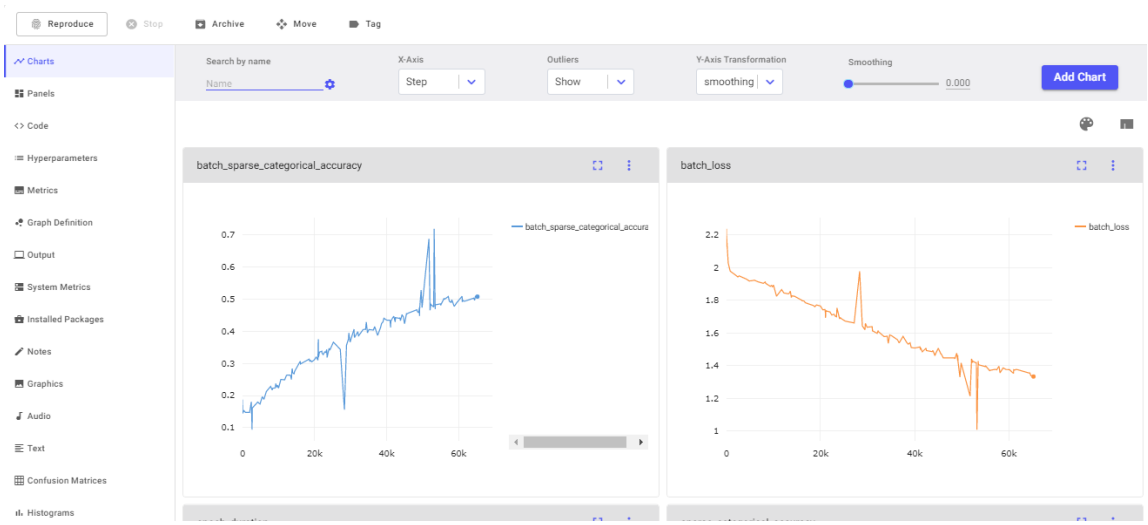


Fig. 3.3. Menú de experimento concreto de CometML con gráficas de métricas generadas

3.3. Requisitos del sistema

Todo sistema que tenga intención de ser utilizado debe tener una serie de requisitos establecidos que permitan establecer cuáles son sus funcionalidades y cómo se realizan. El proceso de establecer los requisitos, aunque se realiza en su mayoría antes del desarrollo del sistema, es iterativo, por lo que según van cambiando las características del proyecto, estos requisitos se pueden adaptar a las nuevas condiciones. Todos ellos deben cumplir una serie de características principales: corrección, coherencia, claridad, consistencia, verificabilidad, comprensibilidad y rastreabilidad.

Para definirlos se utiliza una representación tabular cuya plantilla se muestra en la TABLA. 3.1.

TABLA 3.1. PLANTILLA DE REQUISITOS DEL SISTEMA

Identificador	**RY-XX**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción			

Los atributos se explican a continuación:

- **Identificador:** código único que identifica al requisito. El valor **Y** puede tomar los valores **F** si es funcional y **NF** si no lo es. El valor **XX** es un número de dos cifras que se va incrementando unitariamente.
- **Prioridad:** especifica el nivel de importancia.
- **Necesidad:** especifica si el requisito necesita ser implementado, si es recomendable implementarlo o si es opcional.
- **Estabilidad:** especifica el grado de probabilidad de que el requisito cambie a lo largo del desarrollo del proyecto.
- **Complejidad:** especifica la dificultad de conseguir realizar el requisito.
- **Verificabilidad:** especifica la facilidad con la que se puede comprobar que el requisito se cumple.
- **Descripción:** breve especificación de la condición que establece el requisito sobre el sistema.

3.3.1. Requisitos funcionales

Los requisitos funcionales son aquellos que especifican las funciones que debe realizar el sistema.

TABLA 3.2. REQUISITO RF-01

Identificador	**RF-01**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe recibir como entrada un archivo *.wav*		

TABLA 3.3. REQUISITO RF-02

Identificador	**RF-02**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe eliminar las partes en silencio de la señal recibida		

TABLA 3.4. REQUISITO RF-03

Identificador	**RF-03**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe poder procesar entradas de audio con distinta frecuencia de muestreo		

TABLA 3.5. REQUISITO RF-04

Identificador	**RF-04**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe ser general para cualquier registro del habla		

TABLA 3.6. REQUISITO RF-06

Identificador	**RF-05**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe ser general para cualquier idioma		

TABLA 3.7. REQUISITO RF-06

Identificador	**RF-06**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe procesar señales de audio de cualquier persona del vehículo		

TABLA 3.8. REQUISITO RF-07

Identificador	**RF-07**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe transformar la entrada en un espectrograma para su posterior procesado		

TABLA 3.9. REQUISITO RF-08

Identificador	**RF-08**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema realizará una predicción sobre la entrada procesada y dirá con qué emoción se corresponde		

TABLA 3.10. REQUISITO RF-09

Identificador	**RF-09**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe comunicar al sistema actuador la emoción correspondiente		

3.3.2. Requisitos no funcionales

Los requisitos no funcionales son aquellos que especifican cómo han de realizarse las funcionalidades del sistema previamente definidas en los requisitos funcionales.

TABLA 3.11. REQUISITO RNF-10

Identificador	**RNF-10**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe estar programado en el lenguaje de programación Python		

TABLA 3.12. REQUISITO RNF-11

Identificador	**RNF-11**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe dar una respuesta en tiempo real		

TABLA 3.13. REQUISITO RNF-12

Identificador	**RNF-12**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe estar disponible durante toda la conducción		

TABLA 3.14. REQUISITO RNF-13

Identificador	**RNF-13**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe procesar los audios con la librería Librosa		

TABLA 3.15. REQUISITO RNF-14

Identificador	**RNF-14**		
Prioridad	Baja	Media	Alta
Necesidad	Opcional	Deseable	Esencial
Estabilidad	Baja	Media	Alta
Complejidad	Baja	Media	Alta
Verificabilidad	Baja	Media	Alta
Descripción	El sistema debe realizar las predicciones mediante un modelo de Deep Learning desarrollado con Keras y Tensorflow		

3.4. Casos de uso

Para terminar de definir las funcionalidades que va a realizar el sistema se procede a definir los casos de uso, los cuales mediante el usuario definen las acciones del sistema y como este las utiliza. Debido a que se trata de un sistema automático el usuario no interviene a la hora de utilizar las funcionalidades si no que es el propio sistema el que las realiza. Sin embargo, el sistema no puede funcionar sin el usuario ya que necesita que este hable para que funcione.

Para definirlos se utiliza una representación tabular cuya plantilla se muestra en la TABLA. 3.16.

TABLA 3.16. PLANTILLA CASOS DE USO

Identificador	**CU-XX**
Descripción	
Actores	
Precondiciones	
Postcondiciones	

Los atributos se explican a continuación:

- Identificador: código único que identifica el caso de uso. El valor **XX** es un número de dos cifras que se va incrementando unitariamente.

- Descripción: breve especificación de la finalidad del caso de uso.
- Actores: agentes involucrados en la realización del caso de uso.
- Precondiciones: estado previo que es necesario para que se pueda realizar el caso de uso.
- Postcondiciones: estado posterior en el que se encuentra el sistema tras realizar el caso de uso.

TABLA 3.17. CASO DE USO CU-01

Identificador	**CU-01**
Nombre	Recepción de señal de audio
Descripción	El sistema recibe la señal del audio de un integrante del vehículo para poder procesarla
Actores	Integrante del vehículo
Precondiciones	El sistema se encuentra activo y el usuario habla
Postcondiciones	Se obtiene un archivo de audio listo para ser procesado

TABLA 3.18. CASO DE USO CU-02

Identificador	**CU-02**
Nombre	Procesado del audio
Descripción	El sistema procesa el audio y lo convierte para poder realizar una predicción
Actores	-
Precondiciones	Se ha obtenido un archivo de audio
Postcondiciones	El archivo se ha transformado en un espectrograma para poder procesarlo

TABLA 3.19. CASO DE USO CU-03

Identificador	**CU-03**
Nombre	Predicción de emoción
Descripción	El sistema realiza una predicción de la emoción del usuario a partir del habla
Actores	-
Precondiciones	Se ha procesado el archivo de audio
Postcondiciones	Se obtiene una predicción de una serie de emociones

TABLA 3.20. CASO DE USO CU-04

Identificador	**CU-04**
Nombre	Comunicación al sistema actuador
Descripción	El sistema comunica a un sistema actuador la emoción obtenida
Actores	-
Precondiciones	Se ha realizado una predicción de la emoción
Postcondiciones	El sistema actuador tiene la información necesaria para actuar en consecuencia

3.5. Matriz de trazabilidad entre casos de uso y requisitos

Para asegurar que todas las funcionalidades descritas por los requisitos funcionales han sido contempladas por la utilización del sistema definida por los casos de uso se realiza una matriz donde se vinculan requisitos con los casos de uso que los cubren. Esta matriz se muestra en la TABLA 3.21.

TABLA 3.21. MATRIZ DE TRAZABILIDAD ENTRE REQUISITOS Y CASOS DE USO

Casos de uso / Requisitos	**CU-01**	**CU-02**	**CU-03**	**CU-04**
RF-01	X			
RF-02		X		
RF-03		X		
RF-04			X	
RF-05			X	
RF-06	X			
RF-07		X		
RF-08			X	
RF-09				X

3.6. Diagrama de casos de uso

Los casos de uso pueden representarse mediante un diagrama para que sea más comprensible a simple vista, el cual se muestra en la Fig. 3.4.

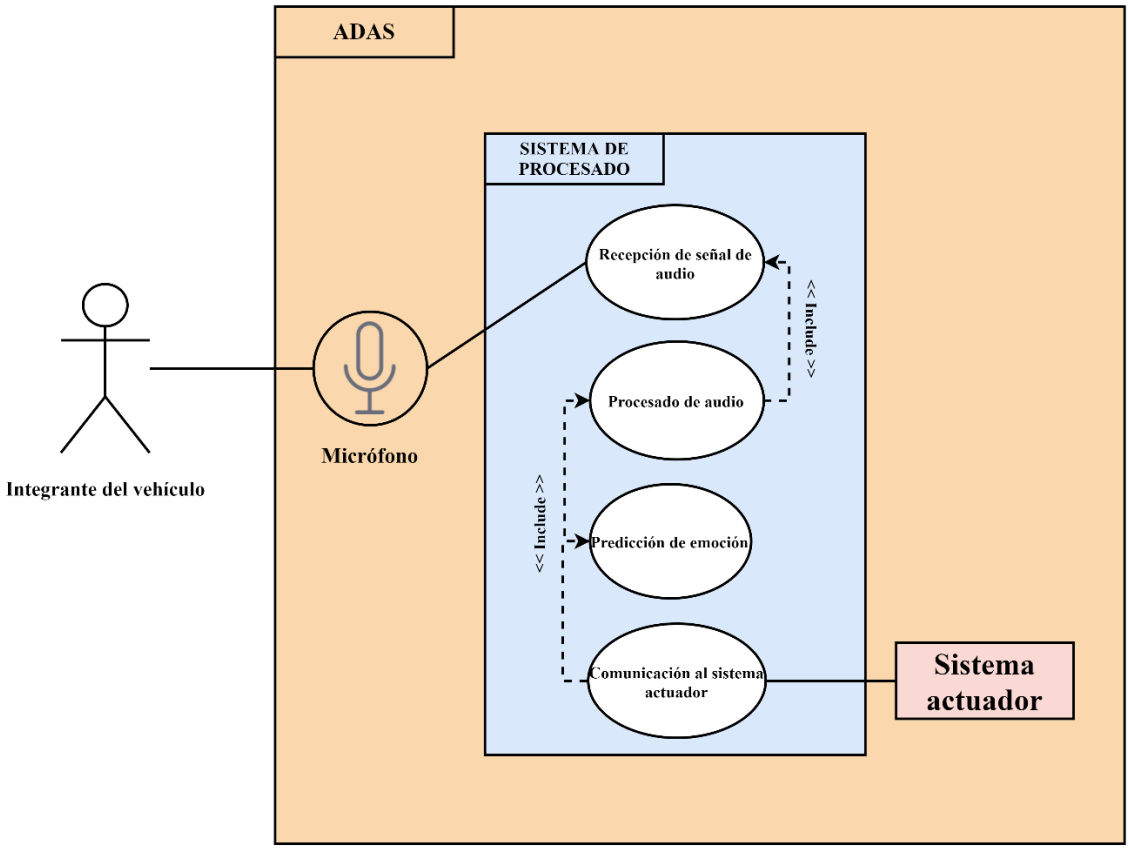


Fig. 3.4. Diagrama de casos de uso

En la parte izquierda del diagrama se encuentra el actor que interactúa con el sistema, el cual es cualquier integrante del vehículo. Este actor no realiza una interacción proactiva con el sistema ADAS si no que es el propio ADAS el cual monitoriza al actor sin que éste intervenga.

El rectángulo situado a la derecha del actor representa el ADAS completo donde se integraría el sistema diseñado, el cual está contenido en el interior del rectángulo del ADAS con el título "*Sistema de procesado*" y que alberga los casos de uso en su interior. Existe una relación entre uno de los casos de uso y un elemento que representa el micrófono del ADAS que a su vez se relaciona con el integrante del vehículo, el cual representa la recogida del audio. Entre los casos de uso existen una serie de relaciones que contienen la etiqueta << *Include* >> la cual significa que hay una relación de dependencia entre un caso de uso que necesita que otro se complete antes. Al tratarse de un sistema secuencial y que solo aporta una funcionalidad, cada caso de uso es dependiente del anterior.

Finalmente hay una relación entre el último caso de uso y el denominado sistema actuador, cuya comunicación contiene la emoción asociada al audio procesado. Esta relación permite al sistema actuador tomar una decisión basada en la emoción detectada de los integrantes del vehículo.

4. OBTENCIÓN Y PROCESADO DE DATOS

Todos los proyectos de minería de datos tienen en común que, a pesar de utilizar técnicas muy diversas y con objetivos muy dispares, necesitan de una fuente de datos para poder funcionar. Estos datos necesitan ser de calidad, variados y equilibrados si se quiere poder realizar un sistema que sea capaz de generalizar un problema y por consiguientemente servir de utilidad.

En este apartado se desarrolla todo lo relativo a los datos utilizados en este proyecto: su obtención, extracción, etiquetado, procesado y los conjuntos de datos generados al final de esta actividad.

4.1. Procedencia y características de los datos

La clasificación de emociones es un campo donde no existe un consenso estricto en el sentido de un grupo de emociones cerradas y absolutas donde se puedan agrupar todos los registros, si no que existen distintos matices o combinaciones de emociones que hacen que la variedad de las clases dentro de los conjuntos de datos sea muy diversa.

Esta diversidad y la dificultad de clasificar emociones debido a sus diferentes manifestaciones dependiendo de cada persona, hacen que en la mayoría de las bases de datos los registros se encuentren etiquetados, es decir, que haya un atributo cuyo valor se corresponde a la emoción que representan el resto de los atributos. Este aspecto influye más adelante en la decisión del tipo de técnicas utilizadas para predecir las emociones, las cuales serán de aprendizaje supervisado.

A continuación, se especifica la procedencia y las características de cada una de las bases de datos que se han utilizado para este proyecto de minería de datos.

4.1.1. Toronto Emotional Speech Set (TESS)

Este dataset generado por la Universidad de Toronto [37] recoge un total de 2800 archivos de audio los cuales recogen el habla de dos actrices de 26 y 64 años y que se encuentran etiquetados mediante la carpeta en la que se encuentran dentro del directorio. Las emociones que se etiquetan son miedo, sorpresa agradable, tristeza, ira, disgusto, felicidad y neutral.

El contenido de los audios se trata de una estructura fija donde cada una de las actrices dice en inglés las palabras "*Say the word X"* donde X es una palabra de dentro de un conjunto de 200 iguales para todas las emociones. Para cada una de las emociones las actrices realizaban una representación artificial del sentimiento en concreto relacionado con ellas.

El formato de los audios es *.wav* y tienen una duración variable dependiendo de la longitud de la palabra concreta que se dice. La frecuencia de muestreo es de 22050 Hz, lo que implica que se toman 22050 valores de la amplitud de la onda cada segundo.

El conjunto de datos TESS tiene la cualidad de ser homogéneo en cuanto a la manera de manifestar las emociones mediante el habla, ya que al haber solamente dos personas que generan los audios, se consigue no tener registros los cuales se ven afectados por la manera de hablar concreta de una persona. Por otro lado, también presenta una serie de desventajas como pueden ser un desbalanceo en aspectos demográficos como la edad, el sexo de las muestras o que el hecho de tener una frase prestablecida puede ofrecer poca información relevante.

4.1.2. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

Este dataset [38] recoge un total de 7356 archivos entre los que se encuentran archivos únicamente de audio, únicamente de video y con audio y video a la vez (en este caso solamente se utilizan los de audio ya que el resto no aporta nada a este proyecto). En ellos 12 actores y 12 actrices vocalizan o cantan dos frases predeterminadas en inglés con distintas emociones y distintas intensidades. Estas emociones son neutral, calma, felicidad, tristeza, ira, miedo, disgusto y sorpresa; las cuales se encuentran codificadas en el nombre del archivo.

Las frases que recitan los actores son "*Kids are talking by the door*" y "*Dogs are sitting by the door*". Los actores tienen un acento norteamericano y realizan dos muestras por frase y emoción ya que utilizan distintos niveles de intensidad de la emoción: alta o baja.

El subconjunto del dataset que contiene a los actores diciendo la frase con una determinada emoción contiene 1140 archivos *.wav*. Estos archivos tienen una frecuencia de muestreo de 48000 Hz y una longitud variable dependiendo de la frase que se dice.

Una característica específica de esta base de datos es que se realizó un proceso de validación de la calidad de los datos una vez se recogieron. En ella un total de 247 voluntarios evaluó en validez, intensidad y genuinidad cada uno de los registros. Este proceso de evaluación concluyó con resultados positivos indicando la calidad y validez del conjunto de datos.

Los puntos a favor de este dataset son el balanceo de las muestras en cuanto a sexo y edad, y la existencia de distintas intensidades de las emociones. Por otro lado, el escaso número de frases que expresan los actores hace que pueda ser poco representativo en ese aspecto.

4.1.3. Berlin Database of Emotional Speech (BERLIN)

Este dataset generado en la Universidad Técnica de Berlín [39] recoge un total de 535 archivos de audio donde 5 actores y 5 actrices de entre 20 y 35 años recitan un total de 10 frases distintas preestablecidas recreando una emoción concreta. El listado de emociones es ira, aburrimiento, disgusto, ansiedad/miedo, felicidad, tristeza y neutralidad; las cuales se encuentran codificadas en el nombre del archivo.

Las frases que dicen los actores no siguen ninguna estructura en concreto, solamente hay cinco de ellas que tienen más extensión que las otras cinco. Estas frases están en alemán y son recitadas en ese idioma por los actores.

Los archivos tienen un formato *.wav* con una frecuencia de muestreo de 22050 Hz. La longitud de los audios es variable dependiendo de si se trata de una de las frases largas o de las cortas.

Las ventajas de este conjunto de datos son su equilibrio demográfico y también la variedad de frases que los actores representan. Entre las desventajas se encuentra el reducido número de archivos que hay comparados con otros *datasets* o la escasez de otros conjuntos de datos con estas características en alemán.

Cabe remarcar que, a pesar de tener ciertos puntos débiles, este dataset es fundamental para poder observar el comportamiento de los modelos a la hora de ser entrenados con distintos idiomas y que efecto se consigue al utilizar solamente uno.

4.2. Extracción y etiquetado

Para poder extraer información de los datos de las distintas fuentes que luego permita al modelo de *Deep Learning* generalizar el problema, es necesario unificar el formato de los registros y también de sus etiquetas. El objetivo de esta fase es convertir las tres bases de datos con una estructura de archivos distinta cada una, en una sola estructura de datos que facilite trabajar con ellos en la fase de procesado de datos y más adelante en el entrenamiento de los modelos.

En un primer lugar, los conjuntos de datos tienen una estructura de archivos donde separan en distintas carpetas las emociones, como es el caso de TESS, o en el que las emociones se distinguen por el nombre del archivo, como sucede en RAVDESS y BERLIN. En el primer caso, las carpetas se acceden una a una y cuando se procesan los archivos dentro de ellas se les asocia la etiqueta respectiva a la carpeta. En el otro caso, la asignación de la etiqueta es individual para cada uno de los archivos que se van procesando.

De cara al proceso de entrenamiento de los modelos y también para ahorrar espacio de almacenaje, las etiquetas son números enteros del 0 al 8. En la TABLA. 4.1 se muestra a que emoción está asociado cada uno de estos valores.

TABLA 4.1. CODIFICACIÓN NUMÉRICA DE LAS EMOCIONES PRESENTES EN LOS DATASETS

Valor numérico	Emoción
0	Ira
1	Disgusto
2	Miedo/Ansiedad
3	Felicidad
4	Neutral
5	Sorpresa
6	Tristeza
7	Aburrimiento
8	Calma

Tras especificar cómo se representan las etiquetas es necesario definir cómo se representan los datos en el conjunto unificado. Todos los archivos de audio de los *datasets* tienen un formato *.wav* al cual va asociado una frecuencia de muestreo y que tienen una duración específica. La función *load()* de la librería Librosa permite transformar un archivo *.wav* en un array de números decimales los cuales representan el valor de la amplitud en un cierto instante de tiempo, como se muestra en la Fig.4.1.

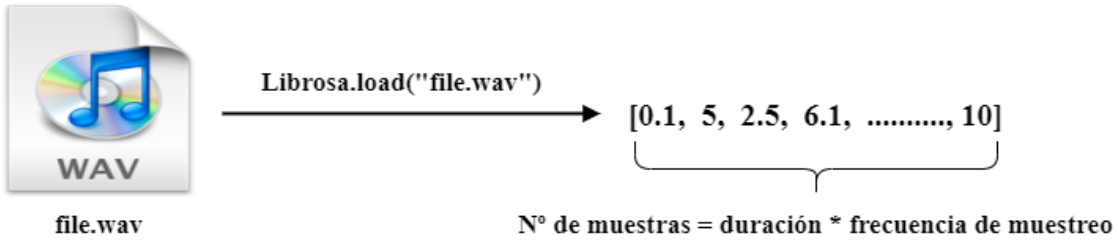


Fig. 4.1. Transformación realizada por la función *load()*

Sin embargo, al tener distintas duraciones y frecuencias de muestreo, el número de elementos de esos arrays es variable, lo que no permite unificarlos. Para solucionar este problema se utiliza la técnica de la segmentación, la cual genera a partir de un registro de mayor tamaño más partes más pequeñas pero iguales en longitud.

El factor utilizado para segmentar los archivos fue de 20.000 muestras. Este valor es debido a que es lo suficientemente grande como para obtener registros con la información necesaria para el entrenamiento de *Deep Learning*, pero también lo bastante pequeño como para no desperdiciar demasiados segmentos de los arrays, ya que cuando al segmentar queda un sobrante que no llega al tamaño de la ventana este se desecha. Este valor genera ventanas de entre 0.9 y 0.4 segundos dependiendo de la frecuencia de muestreo.

A parte de la segmentación, en este proceso de extracción se realiza un proceso de recorte de las partes del audio que se encuentran en silencio. Esto se realiza debido a que, en los experimentos que se realizaron para obtener los *datasets*, se grabaron los audios de tal manera que el instante en el que el sujeto empieza a hablar no es justo la primera muestra que tomó el micrófono, si no que pasan unas ciertas décimas de segundo. Estas muestras con valor nulo sumadas a aquellos silencios que hace una persona entre palabra y palabra crean una gran cantidad de atributos con valores nulos que pueden afectar al aprendizaje de las redes de neuronas.

Para realizar el proceso de eliminación de las partes en silencio se utiliza una función de la librería Librosa llamada *trim()*. La sensibilidad de la función puede ser regulada con los parámetros para que elimine todas las partes con sonido, pero no se aumenta esta sensibilidad, ya que al hacerlo hay ciertas frecuencias e intensidades que son también eliminadas y por lo tanto se pierde información. En la Fig. 4.2 se muestra el resultado de aplicar esta función a una pista de sonido, de la cual se eliminan 0.5 segundos de silencio.

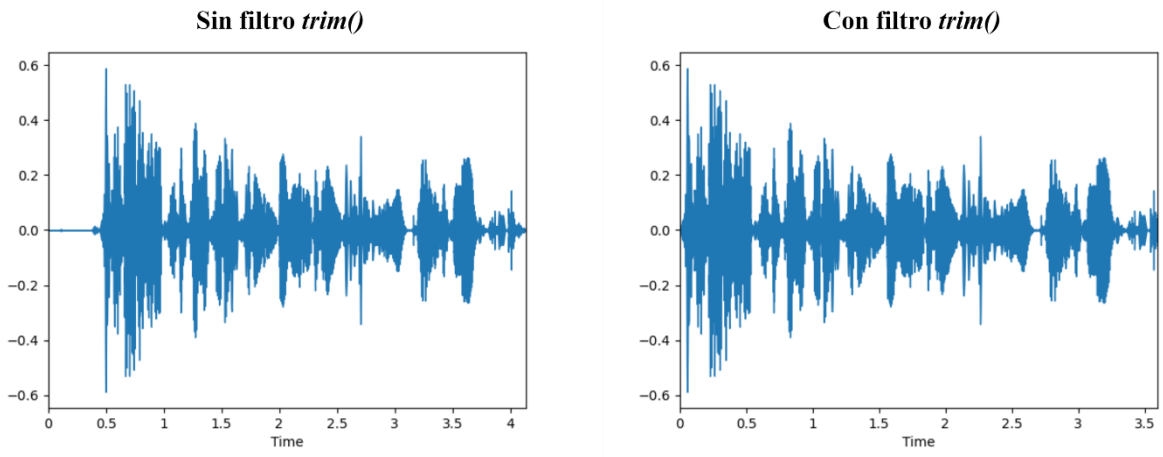


Fig. 4.2. Pista de sonido antes y después de aplicar la función *trim()*

Tras estas transformaciones se obtiene un conjunto de datos los cuales tienen el mismo número de atributos y un conjunto de etiquetas homogéneas, lo cual permite realizar el procesado para obtener características ricas en información.

4.3. Procesado de datos

Una vez se obtiene un conjunto de datos en el que se puede trabajar, es necesario mediante la aplicación de filtros y procesos, enriquecer y seleccionar aquellos atributos que aportan información de utilidad para la tarea de la clasificación.

Tras el proceso de extracción y etiquetado, los registros tienen estructura de array unidimensional en la cual se representa la amplitud respecto al tiempo. Esta representación de una onda de sonido es en verdad una simplificación, ya que a la hora de digitalizar el sonido es imposible recoger una serie de valores continuos a no ser que

sea expresándolo mediante una función, por lo que se realiza un muestreo en instantes concretos.

La onda de sonido que se transmite por el aire rara vez es una onda pura, la cual se caracteriza por tener una frecuencia constante y una amplitud variable con respecto al tiempo como se muestra en la Fig. 4.3, si no que se trata de la superposición de distintas ondas puras cuya agregación crea una onda compleja.

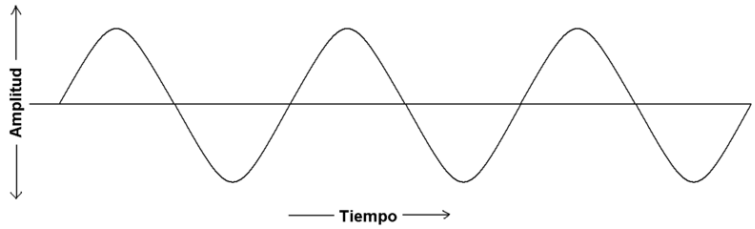


Fig. 4.3. Representación de onda sinusoidal pura [40]

Existe una transformación matemática denominada la Transformada de Fourier que es capaz de, a partir de una onda compleja en el dominio del tiempo, obtener las frecuencias puras que la componen mediante su agregación. Este proceso se muestra de manera gráfica en la Fig. 4.4, donde una onda compleja se descompone en cuatro ondas puras más simples.

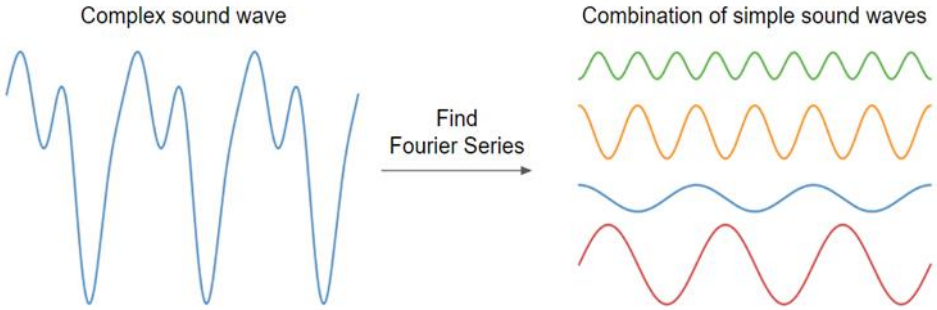


Fig. 4.4. Proceso de descomposición de la Transformada de Fourier [41]

La Transformada de Fourier es relevante ya que es el núcleo de un procedimiento denominado Coeficientes Cepstrales de las Frecuencias de Mel (MFCC), el cual se utiliza para obtener una representación más rica en información, ya que aporta una representación más detallada de las frecuencias de la onda, y que se utiliza para entrenar las redes de neuronas convolucionales.

El procedimiento de MFCC consta de una serie de pasos que transforman el array de una dimensión en una matriz de dos dimensiones:

1. Segmentar la entrada en ventanas de tiempo que se superponen.

2. Aplicar la Transformada de Fourier a cada una de esas ventanas para así obtener el valor de las frecuencias puras que la forman y la intensidad de cada una.

3. Aplicar la escala de Mel a cada una de las frecuencias obtenidas en el paso anterior.

4. Tomar el logaritmo de cada uno de los coeficientes.

5. Aplicar la transformada del coseno discreta a cada uno de los resultados del paso anterior.

Este procedimiento aplica en el paso número tres una transformación de las frecuencias que se convierten en una magnitud de la escala de Mel. Esta escala musical es de gran importancia debido a que, mediante una transformación no lineal, transforma las frecuencias de hercios a mels, los cuales son unidades que van marcadas por la capacidad que tiene un ser humano de diferenciar distintas frecuencias. En la Fig. 4.5 se observa la relación entre unidades de hercios y mels, habiendo una gran diferencia entre los mels que equivalen a valores de hercios bajos, debido a que el oído humano es capaz de diferenciarlos de manera precisa, mientras que apenas cambian los mels para los valores altos ya que son imperceptibles para las personas.

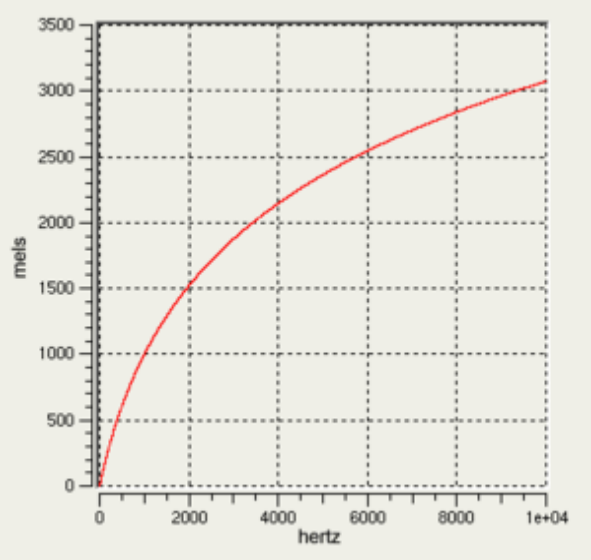


Fig. 4.5. Función de la transformación no lineal de hercios a mels [42]

Aplicado a la práctica, la transformación de MFCC se utiliza mediante la función *filters.mfcc()* de la librería Librosa la cual recibe los siguientes argumentos:

- Array tras el procesado del archivo *.wav*.
- Frecuencia de muestreo del archivo.
- Tamaño de las ventanas temporales que se especifican en el paso uno del procedimiento.

- Desplazamiento de las ventanas a través del array.
- Número de mels que se utilizan para crear la matriz de la salida.

La aplicación de este proceso, el cual se muestra en la Fig. 4.6 convierte los registros de una a dos dimensiones, lo que conlleva ganar una gran cantidad de información a parte de obtener un formato más adecuado para el uso de redes de neuronas convolucionales.

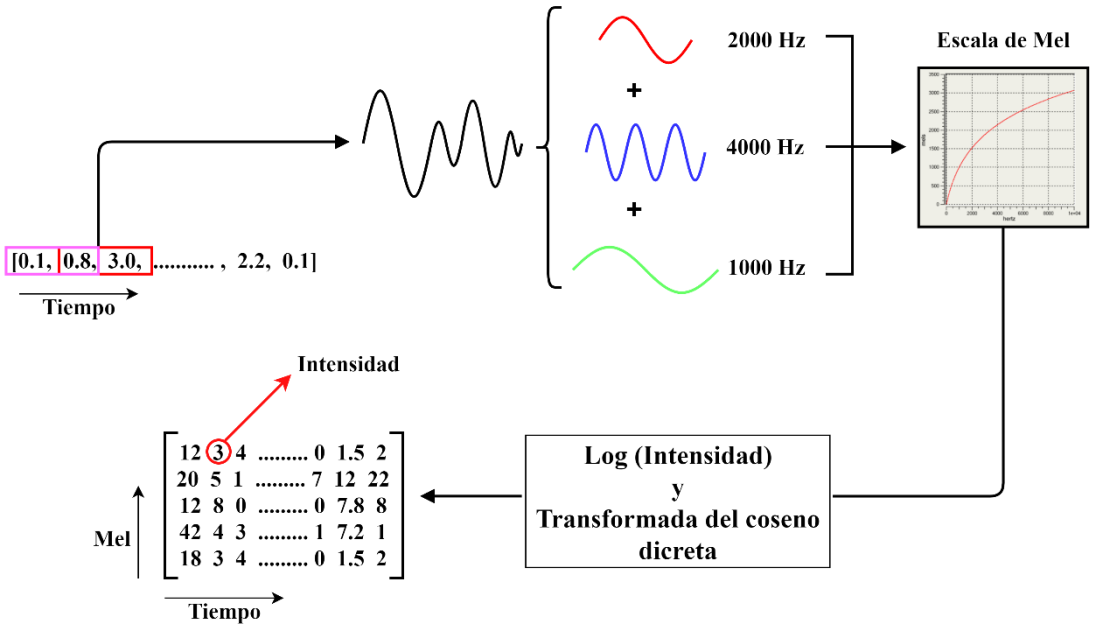


Fig. 4.6. Diagrama del proceso de la transformación MFCC

4.4. Generación de los conjuntos de datos

Una vez se obtienen los registros procesados de las distintas fuentes de datos, es necesario definir una o varias combinaciones de estos dataset para utilizar en el proceso de aprendizaje de las redes de neuronas. Este proceso se realiza ya que los resultados del aprendizaje no solamente dependen de la técnica de aprendizaje automático utilizada y los parámetros de esta, si no que el conjunto de datos tiene una gran influencia sobre los resultados también.

Con la finalidad de abordar el problema desde diversas metodologías, aparte de utilizar las matrices de números decimales, también se crea a partir de ellas un conjunto de datos que utiliza una representación distinta: imágenes.

Este conjunto se genera mediante la representación gráfica de los valores de las matrices obtenidas en el procesado de datos en forma de espectrograma. Este tipo de representación, como se puede observar en la Fig. 4.7, muestra en los ejes de tiempo y frecuencia los valores de intensidad que están representados con distintos colores dependiendo de su valencia.

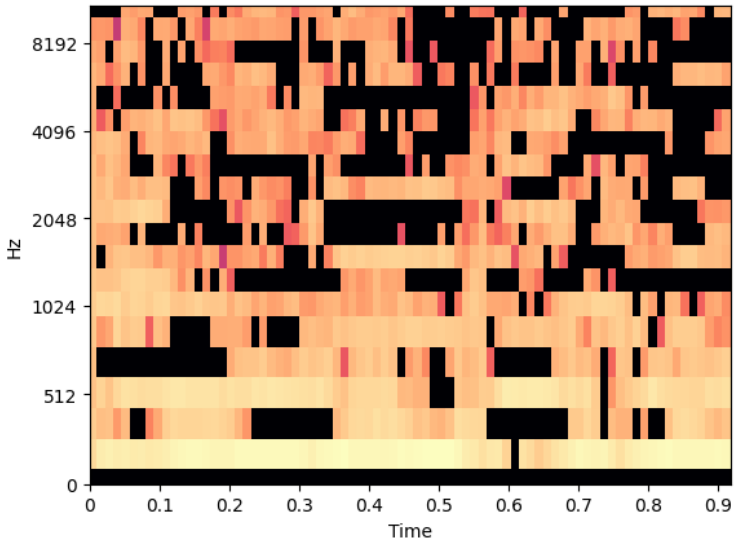


Fig. 4.7. Representación en formato de espectrograma de la transformación MFCC de un registro de audio

La razón de aplicar esta transformación a imágenes es la posibilidad de generar instancias con más información que las matrices de valores. Esta ganancia de información se debe a dos razones principales:

- La representación de las ventanas de tiempo no aporta ninguna información en el caso de representación con matrices, mientras que, a la hora de transformarlo a imágenes, el tamaño de los rectángulos que representan esas ventanas de tiempo es directamente proporcional a su duración y su desplazamiento como se muestra en la Fig. 4.8 donde se comparan dos representaciones del mismo audio con tamaños de ventanas y desplazamiento distintos.

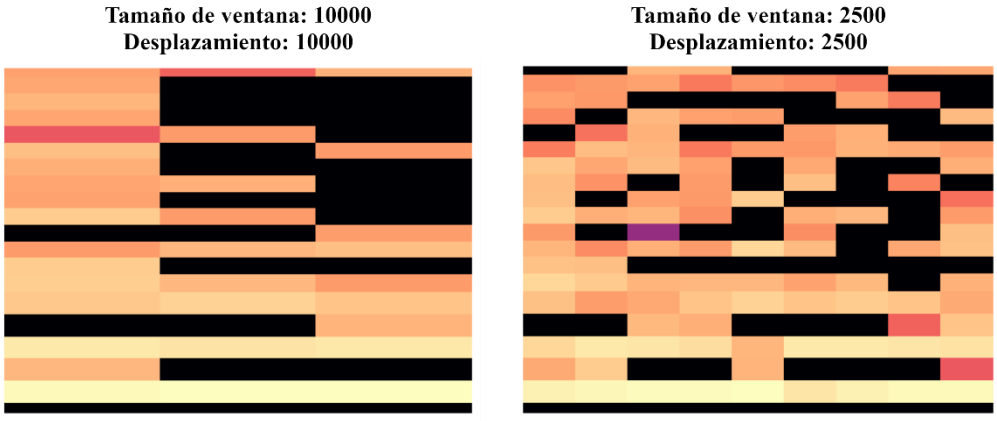


Fig. 4.8. Comparación entre las imágenes generadas con distinto tamaño de ventana y desplazamiento

- La representación de los mels es distinta ya que, a la hora de representarlo en imágenes, el eje vertical se encuentra en hercios. Esto provoca que intensidades de valores de mels más bajos tengan un tamaño vertical mayor y las intensidades de mayor valor de mels tengan un tamaño menor, debido a la

relación no lineal entre mels y hercios explicada previamente. Este fenómeno se muestra en la Fig. 4.9 donde la altura de los rectángulos superiores es menor que la de los inferiores.

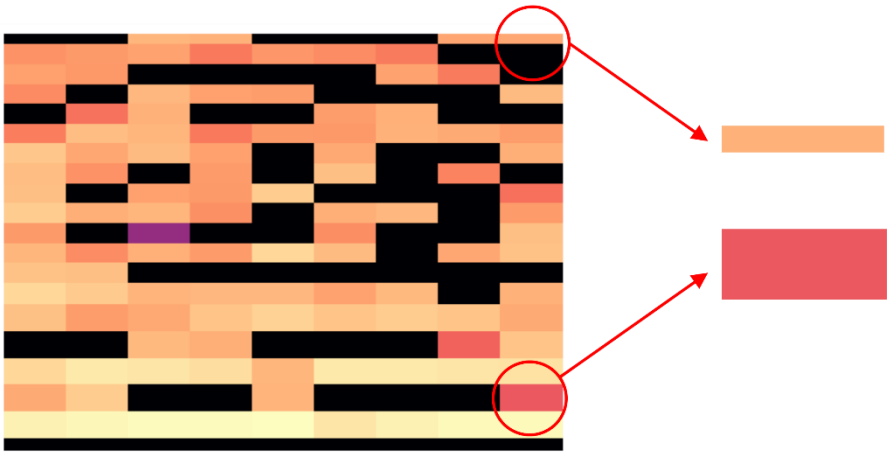


Fig. 4.9. Comparación del tamaño entre ventanas de frecuencias altas con ventanas de menor altura (naraja) y frecuencias bajas con mayor altura vertical (rojo)

A parte de generar una manera diferente de expresar las instancias, también se combinan las distintas fuentes de datos y se crean subconjuntos de ellas para obtener el conjunto de bases de datos de entrenamiento detallado en la TABLA 4.2.

TABLA 4.2. SUBCONJUNTOS DE DATOS GENERADOS

Código	***Datasets***	**Descripción**
RBT_I	RAVDESS, BERLIN, TESS	*Dataset* de imágenes de todo el conjunto de datos
RBT_M	RAVDESS, BERLIN, TESS	*Dataset* de matrices de valores decimales de todo el conjunto de datos
RT_I	RAVDESS, TESS	*Dataset* de imágenes sin el conjunto BERLIN
RT_M	RAVDESS, TESS	*Dataset* de matrices de valores decimales sin el conjunto BERLIN
RB_I	RAVDESS, BERLIN	*Dataset* de imágenes sin el conjunto TESS
RB_M	RAVDESS, BERLIN	*Dataset* de matrices de valores decimales sin el conjunto TESS

Una vez creados todos los subconjuntos de datos, se realiza un balanceo de las clases de cada uno de ellos. Este proceso de balanceo consiste en modificar las instancias del conjunto de datos con el objetivo de equilibrar la representación de cada uno de los casos que se quiere clasificar en el problema. Este desbalance proviene de mezclar *datasets*

distintos que tienen distintas clases y cuyas instancias contienen características específicas. Por ejemplo, al combinar el conjunto de datos RAVDESS con cualquiera de los otros dos la clase *calma* está desbalanceada, ya que RAVDESS es el único dataset que la tiene y al combinarse con los otros queda con muchas menos instancias con respecto a las otras clases las cuales se suman.

Existen diversas técnicas de balanceo como pueden ser el submuestreo, sobremuestreo, adaptación de métricas de evaluación, etc. En este caso, se utiliza el submuestreo, el cual consiste en eliminar aleatoriamente instancias de las clases que más tengan hasta igualar en número a las menos representativas. Se selecciona esta técnica ya que disponiendo de un conjunto de datos lo suficientemente amplio, es la aproximación más sencilla de implementar y a la vez aquella que menos sesgo introduce en el conjunto de datos.

Hay casos en los que realizar submuestreo reduciría demasiado el número de instancias de las distintas clases ya que la clase menos representativa tienen significativamente menos registros que las demás. En este caso la solución utilizada es en primer lugar intentar combinar esa clase con alguna otra clase equivalente y en caso contrario eliminar la clase por completo.

Tras realizar el proceso de balanceo, los subconjuntos definidos anteriormente en la TABLA 4.2 tienen las características definidas en las TABLA 4.3 y TABLA 4.4.

TABLA 4.3. DEFINICIÓN DE LAS CARACTERÍSTICAS DE LOS SUBCONJUNTOS DE DATOS (PARTE1)

Dataset	**Número de instancias**	**Ajustes realizados**
RBT_I	8750	Se eliminan las clases *aburrimiento* y *calma*. Se realiza un submuestreo de 1250 instancias por clase.
RBT_M	8750	Se eliminan las clases *aburrimiento* y *calma*. Se realiza un submuestreo de 1250 instancias por clase.
RT_I	7700	Se elimina la clase *aburrimiento*. Se realiza un submuestreo de 1100 instancias por clase
RT_M	7000	Se elimina la clase *aburrimiento*. Se realiza un submuestreo de 1100 instancias por clase

TABLA 4.4. DEFINICIÓN DE LAS CARACTERÍSTICAS DE LOS SUBCONJUNTOS DE DATOS (PARTE 2)

RB_I	5040	Se combinan las clases *neutral* y *aburrimiento*. Se realiza un submuestreo de 630 instancias por clase
RB_M	5040	Se combinan las clases *neutral* y *aburrimiento*. Se realiza un submuestreo de 630 instancias por clase

En ciertos casos las clases *neutral, aburrimiento* o *calma* se eliminan o se combinan entre ellas para balancear el conjunto. Estos cambios se pueden realizar debido a que las clases son equivalentes o no tienen relevancia dentro del contexto del problema que se quiere abordar, que es el de detectar emociones que pueden afectar a la conducción.

Cabe aclarar que el desarrollo de estos subconjuntos de datos se obtiene en el transcurso de la experimentación, donde se realizan cambios orientados por los resultados obtenidos. Los dos cambios principales realizados tienen que ver con la eliminación de uno de los tres subconjuntos que forman la totalidad de la base de datos recopilada.

La eliminación del dataset BERLIN en los conjuntos *RT_I* y *RT_M,* se realiza con la finalidad de obtener una mejoría en los resultados al eliminar el idioma minoritario de los registros, el cual podría estar afectando a los resultados por las diferencias fonéticas remarcables del alemán respecto al inglés del resto de datos.

La eliminación del dataset TESS en los conjuntos *RB_I* y *RB_M,* se realiza con la finalidad de obtener una mejoría en los resultados al eliminar el desequilibrio de muestras femeninas frente a las masculinas, ya que debido a factores biológicos hay ocasiones en las cuales la diferencia en las frecuencias tiene valores significativos que pueden afectar a la capacidad de generalización si el conjunto de datos no se encuentra balanceado.

5. RECONOCIMIENTO DE EMOCIONES

En este apartado se describe el proceso de creación de los modelos encargados de realizar el reconocimiento de la emoción del conductor a partir del habla. La obtención de estos se realiza mediante un proceso iterativo donde se prueban distintas configuraciones y *datasets* mientras se evalúa su rendimiento. El producto final de este proceso consiste en aquel modelo que realiza mejor la tarea de generalizar la clasificación de las entradas y que por lo tanto va a tener un mejor desempeño en un entorno real. En este proceso de experimentación intervienen múltiples factores los cuales se discuten a continuación.

5.1. Hiperparámetros del proceso de experimentación

Un factor esencial en la creación de un modelo basado en redes de neuronas es la decisión de una serie de hiperparámetros, los cuales son valores no entrenables que especifica el diseñador del modelo, y que influyen en su resultado y en cómo se llega a él.

Los hiperparámetros de las redes de neuronas se pueden dividir en dos grupos: los relativos a la topología de la red y los relacionados con el proceso de aprendizaje. Debido al gran número de hiperparámetros respectivos a la topología, estos se especifican en el siguiente apartado, pero los relativos al aprendizaje se detallan sus valores en la TABLA 5.1 y se explican a continuación.

TABLA 5.1. VALOR Y DESCRIPCIÓN DE LOS HIPERPARÁMETROS DEL PROCESO DE EXPERIMENTACIÓN

Nombre	Valor	Descripción
Tasa de aprendizaje	Variable	Su valor se obtiene mediante la experimentación y depende del resto de hiperparámetros
Nº de *epochs*	Variable	Su valor se obtiene mediante la experimentación y depende principalmente del valor de la tasa de aprendizaje
Optimizador	*Adamax* o *SGD*	Se obtienen mediante la experimentación y comparando cuales dan mejores resultados
Batch size	Variable	Se obtiene mediante la experimentación y depende principalmente del conjunto de datos y el hardware utilizado
Función de error	*Sparse Categorical Cross Entropy*	Se obtiene a partir del tipo de problema abordado

La tasa de aprendizaje es el factor por el que se multiplican los incrementos que se añaden a los pesos durante el proceso de aprendizaje. Su valor influye en la velocidad con la que aprenden las redes. Si su valor es demasiado bajo, la red puede tardar demasiado y requerir muchos ciclos en converger a una solución óptima. Por otro lado, si este valor es demasiado elevado, puede provocar que el aprendizaje oscile en torno a un punto perteneciente a un mínimo local y que por lo tanto se estanque en una solución subóptima. Su valor es obtenido mediante la experimentación y es dependiente del valor del número de ciclos o *epochs*, obteniendo los mejores resultados con valores de 0.0001, 0.0002 y 0.0005 en este caso.

El número de ciclos de aprendizaje o *epochs* es el número de veces que se introduce la totalidad del conjunto de valores de entrenamiento en la red durante el aprendizaje. Este valor es dependiente de la tasa de aprendizaje ya que cuanto menor sea mayor número de *epochs* serán necesarios. Su valor debe ser el suficiente como para asegurar que la red converja hacia una solución óptima, pero sin ser demasiado grande ya que puede provocar sobreaprendizaje. Los valores utilizados que mejores soluciones aportan son aquellos entre 70 y 150 dependiendo de la tasa de aprendizaje.

El optimizador es el algoritmo utilizado para ajustar los parámetros entrenables de la red de neuronas. Su finalidad es minimizar el valor de la función de error para una serie de parámetros, los cuales se corresponden con los pesos de la red. La biblioteca Keras

proporciona una serie de optimizadores, los cuales se prueban de manera experimental para comparar cual es que proporciona mejores resultados. Aquellos que proporcionan mejores resultados son los optimizadores *Adamax* y *SGD*.

El *batch size* determina el número de instancias que tiene un lote de entrenamiento. Este tamaño influye en los resultados y la velocidad del entrenamiento ya que, en un *epoch*, el batch size determina cuantos datos se introducen en la red antes de realizar una actualización de los pesos. Esto significa que por ejemplo un entrenamiento con 100 datos de entrenamiento y un *batch size* de 5, los pesos se actualizan 20 veces por *epoch*. Aumentar su tamaño acelera el tiempo de entrenamiento ya que significa menos actualizaciones y además permite la paralelización de los cálculos; sin embargo, esto puede afectar negativamente a la calidad del aprendizaje ya que puede reducir demasiado el número de actualizaciones.

La función de error se encarga de calcular la desviación del modelo y es aquella métrica que se intenta minimizar en el aprendizaje. Existen gran cantidad de funciones de error las cuales se utilizan dependiendo del tipo de problema y la manera de representar las predicciones. En este caso, ya que se trata de una clasificación de más de dos clases donde las etiquetas tienen valores de números enteros, se utiliza la función *Sparse Categorical Cross Entropy.*

5.2. Arquitectura de la red

En este apartado se especifican las características de las redes de neuronas que se utilizan para generalizar el problema. El punto de partida es la utilización de redes de neuronas convolucionales, las cuales han demostrado un gran rendimiento a la hora de resolver problemas con imágenes y matrices, lo cual las hace recomendables para este proyecto. Otra información previa a construir ningún modelo es la aportada por trabajos similares, como los definidos en los apartados 2.6.1 y 2.6.2, en los cuales a la hora de trabajar con sonido procesado por MFCC se especifica que las redes con topologías más sencillas tienen mejor rendimiento que aquellas que tienen gran cantidad de filtros, neuronas y capas.

Debido a que los subconjuntos de datos de matrices e imágenes necesitan de arquitecturas distintas, se separa en dos apartados a continuación la definición de las topologías.

A pesar de las diferencias, la función de activación de las neuronas es un parámetro que es común a ambas aproximaciones y que es necesario especificar. Existen multitud de funciones de activación las cuales aportan distintos comportamientos a las redes de neuronas. La opción elegida fue la función *Rectified Linear Unit (ReLU)* ya que ha sido demostrada su mejora en rendimiento para problemas de *Deep Learning* frente a otras funciones de activación y desde su aparición en 2010 ha sido ampliamente usada en el estado del arte [43].

5.2.1. Imágenes

En la TABLA 5.2 se detallan las arquitecturas obtenidas en el proceso de experimentación, donde se utilizan distintos códigos de color para cada uno de los tipos distintos de capas. La leyenda de los códigos de color se puede observar en la Fig. 5.1.

TABLA 5.2. ARQUITECTURAS DE LAS CNN UTILIZADAS PARA IMÁGENES

CNN 1	CNN 2	CNN 3	CNN 4
BATCH NORMALIZATION	BATCH NORMALIZATION	BATCH NORMALIZATION	BATCH NORMALIZATION
15 CONV 2D	20 CONV 2D	20 CONV 2D	20 CONV 2D
MaxPooling (4,4)	MaxPooling (2,2)	MaxPooling (2,2)	MaxPooling (2,2)
DROPOUT (0.25)	DROPOUT (0.25)	DROPOUT (0.25)	DROPOUT (0.25)
10 CONV 2D	20 CONV 2D	20 CONV 2D	20 CONV 2D
MaxPooling (2,2)	MaxPooling (2,2)	MaxPooling (2,2)	MaxPooling (2,2)
DROPOUT (0.25)	DROPOUT (0.25)	DROPOUT (0.25)	DROPOUT (0.25)
10 CONV 2D	10 CONV 2D	10 CONV 2D	10 CONV 2D
DROPOUT (0.25)	DROPOUT (0.25)	DROPOUT (0.25)	DROPOUT (0.25)
FLATTEN	FLATTEN	FLATTEN	FLATTEN
DENSE 21	DENSE 70	BATCH NORMALIZATION	BATCH NORMALIZATION
DROPOUT (0.25)	DROPOUT (0.25)	DENSE 70	DENSE 48
DENSE 21	DENSE 21	DROPOUT (0.25)	DROPOUT (0.25)
OUTPUT SOFTMAX	OUTPUT SOFTMAX	DENSE 21	DENSE 21
		OUTPUT SOFTMAX	DROPOUT (0.25)
			DENSE 21
			OUTPUT SOFTMAX

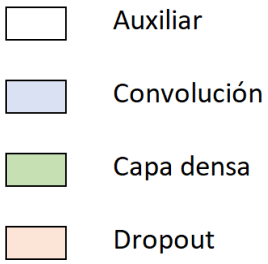


Fig. 5.1. Leyenda de colores de los distintos tipos de capas en las topologías de red

5.2.2. Matrices de valores decimales

En la TABLA 5.3 se muestran las arquitecturas obtenidas en el proceso de experimentación para el problema de clasificación de emociones mediante matrices de valores decimales. Esta tabla continua con la numeración de las arquitecturas y utiliza el

mismo código de color que la mostrada para las imágenes, por lo tanto, utiliza la misma leyenda para los códigos de colores, los cuales se muestran en la Fig. 5.1.

TABLA 5.3. ARQUITECTURAS DE LAS CNN UTILIZADAS PARA MATRICES DE VALORES DECIMALES

CNN 5	CNN 6	CNN 7	CNN 8
BATCH NORMALIZATION	BATCH NORMALIZATION	BATCH NORMALIZATION	BATCH NORMALIZATION
20 CONV 2D	20 CONV 2D	20 CONV 2D	20 CONV 2D
20 CONV 2D	20 CONV 2D	20 CONV 2D	MaxPooling (2,2)
DROPOUT (0.25)	DROPOUT (0.25)	DROPOUT (0.25)	DROPOUT (0.25)
MaxPooling (3,3)	MaxPooling (3,3)	MaxPooling (2,2)	20 CONV 2D
5 CONV 2D	5 CONV 2D	DROPOUT (0.3)	DROPOUT (0.25)
DROPOUT (0.3)	DROPOUT (0.3)	FLATTEN	MaxPooling (2,2)
FLATTEN	FLATTEN	BATCH NORMALIZATION	5 CONV 2D
DENSE 21	BATCH NORMALIZATION	DENSE 21	DROPOUT (0.3)
DROPOUT (0.25)	DENSE 42	DROPOUT (0.25)	FLATTEN
OUTPUT SOFTMAX	DROPOUT (0.25)	DENSE 21	BATCH NORMALIZATION
	OUTPUT SOFTMAX	OUTPUT SOFTMAX	DENSE 70
			DROPOUT (0.25)
			DENSE 21
			OUTPUT SOFTMAX

5.3. Resultados de la experimentación

Este apartado muestra los resultados de los experimentos realizados con los subconjuntos de datos, hiperparámetros y arquitecturas definidas previamente. El objetivo es comparar su desempeño a la hora de generalizar el problema planteado para así poder seleccionar un modelo final que será utilizado en la práctica.

En las TABLA 5.4 y TABLA 5.5 se puede observar como se han realizado un total de 58 experimentos en total. Estos experimentos están definidos por un identificador único, el dataset utilizado para entrenar y validar la red, el optimizador, la tasa de aprendizaje, el tamaño del *batch* y el número de *epochs*.

Cada experimento también va acompañado de la medida de la *accuracy* del conjunto de test. Esta medida expresa el porcentaje de instancias clasificadas correctamente por la red y se calcula dividiendo el número de instancias correctamente clasificadas entre el número total de instancias clasificadas. Para obtener este valor se utiliza el conjunto de test ya que es el conjunto que expresa el grado de generalización de la red a instancias del problema que aún no ha aprendido.

En algunos de los experimentos realizados, aparecen las siglas CV al lado del valor de la *accuracy*, las cuales hacen referencia al término *Cross Validation*, o validación cruzada en castellano. La validación cruzada es una técnica utilizada en la experimentación con redes de neuronas, que consiste en entrenar el modelo varias veces con distintas particiones de los conjuntos de train y test y obtener medidas a partir de la media de esos experimentos. Esta técnica pretende comprobar la validez de los resultados de los experimentos y que la manera de dividir los datos en los distintos conjuntos no introduce ningún sesgo en ellos. Se realiza validación cruzada con tres conjuntos en algunos de los experimentos más relevantes, ya que al no apreciar un cambio significativo no se realizó para otros con menos relevancia, y que aporta la certeza de que los resultados de los experimentos utilizados para comparar aproximaciones no contienen sesgos debido a la división de los conjuntos de entrenamiento y test.

TABLA 5.4. RESULTADOS DE LA EXPERIMENTACIÓN (PARTE 1)

ID	Dataset	Arquitectura	Optimizador	T. Aprend.	Epochs	Accuracy (%)
1	RBT_I	CNN 1	SGD	0.0001	100	51.02
2	RBT_I	CNN 1	SGD	0.0001	200	56.12
3	RBT_I	CNN 2	SGD	0.0005	200	51.57
4	RBT_I	CNN 2	Adamax	0.0005	100	50.31
5	RBT_I	CNN 2	Adamax	0.0002	100	49.09
6	RBT_I	CNN 3	Adamax	0.0002	50	47.74
7	RBT_I	CNN 3	Adamax	0.0002	50	29.89
8	RBT_I	CNN 3	SGD	0.0001	70	55.48
9	RBT_I	CNN 4	Adamax	0.0002	80	47.59
10	RBT_I	CNN 4	Adamax	0.0001	80	40.91
11	RBT_I	CNN 4	SGD	0.0001	80	48.78 CV
12	RT_I	CNN 1	SGD	0.0001	200	91.63
13	RT_I	CNN 1	Adamax	0.0001	100	89.31
14	RT_I	CNN 1	Adamax	0.0001	170	90.47
15	RT_I	CNN 2	Adamax	0.0001	170	88.8
16	RT_I	CNN 2	SGD	0.0001	100	91.24
17	**RT_I**	**CNN 3**	**SGD**	**0.0001**	**100**	**93.34 CV**
18	RT_I	CNN 3	Adamax	0.0001	100	89.31
19	RT_I	CNN 3	Adamax	0.00005	100	90.99
20	RT_I	CNN 4	Adamax	0.0001	45	70.14
21	RT_I	CNN 4	SGD	0.0001	45	86.35
22	RT_I	CNN 4	SGD	0.0001	45	90.47
23	RB_I	CNN 1	Adamax	0.0001	60	21.59
24	RB_I	CNN 1	SGD	0.0001	100	28.61
25	RB_I	CNN 2	SGD	0.0001	100	27.47
26	RB_I	CNN 2	Adamax	0.0001	100	22.62
27	RB_I	CNN 3	SGD	0.0001	100	27.37
28	RB_I	CNN 3	SGD	0.00005	50	28.3
29	RB_I	CNN 3	Adamax	0.00005	50	24.89
30	RB_I	CNN 3	Adamax	0.00005	100	23.03
31	RB_I	CNN 4	SGD	0.0001	100	27.27
32	RB_I	CNN 4	Adamax	0.0001	100	21.69
33	RBT_M	CNN 5	Adamax	0.00005	250	63.64
34	RBT_M	CNN 5	SGD	0.00008	250	63.57

TABLA 5.5. RESULTADOS DE LA EXPERIMENTACIÓN (PARTE 2)

35	RBT_M	CNN 5	SGD	0.0001	250	65.3
36	RBT_M	CNN 6	SGD	0.0001	250	65.16
37	RBT_M	CNN 6	Adamax	0.0001	250	64.16
38	RBT_M	CNN 6	SGD	0.00008	300	66.03
39	RBT_M	CNN 7	SGD	0.0001	250	67.52
40	RBT_M	CNN 7	Adamax	0.0001	250	66.89
41	RBT_M	CNN 8	SGD	0.0001	300	67.41
42	RBT_M	CNN 8	Adamax	0.0001	300	67.40 CV
43	RT_M	CNN 5	Adamax	0.002	250	71.27 CV
44	RT_M	CNN 5	SGD	0.00008	250	70.56
45	RT_M	CNN 6	Adamax	0.0001	300	67.41
46	RT_M	CNN 6	SGD	0.00008	300	65.32
47	RT_M	CNN 7	Adamax	0.0001	250	69.61
48	RT_M	CNN 7	SGD	0.00008	250	68.63
49	RT_M	CNN 8	Adamax	0.0001	300	69.26
50	RT_M	CNN 8	SGD	0.00008	300	68.67
51	RB_M	CNN 5	Adamax	0.002	250	35.81
52	RB_M	CNN 5	SGD	0.00008	250	35.51 CV
53	RB_M	CNN 6	Adamax	0.0001	300	32.89
54	RB_M	CNN 6	SGD	0.00008	300	33.13
55	RB_M	CNN 7	Adamax	0.0001	250	34.26
56	RB_M	CNN 7	SGD	0.00008	250	33.78
57	RB_M	CNN 8	Adamax	0.0001	300	34.86
58	RB_M	CNN 8	SGD	0.00008	300	24.27

Los datos acerca del proceso de experimentación se muestran en la TABLA 5.6, donde se puede apreciar datos como la duración, la cual es muy superior en los *datasets* de imágenes debido a su mayor tamaño, y los resultados generales obtenidos, los cuales tienen un mejor promedio para los *datasets* de matrices de valores decimales pero el caso de mayor precisión se encuentra en los modelos que utilizan imágenes.

TABLA 5.6. RESUMEN DEL PROCESO DE EXPERIMENTACIÓN

Imágenes			**Matrices de valores decimales**		
Tiempo de experimentación	**Resultados**		**Tiempo de experimentación**	**Resultados**	
	Máximo	**Media**		**Máximo**	**Media**
15 horas y 50 minutos	93.34 %	55 %	1 hora	71.27 %	56.8 %

A parte de los experimentos definidos, también se realizan una serie de experimentos que no se incluyen debido a que sus resultados no son válidos debido a errores en los *datasets* utilizados para entrenarlos. Aunque estos experimentos no se muestren, estos toman un papel importante en el proceso de experimentación ya que permiten explorar aquellas arquitecturas e hiperparámetros que ofrecen buenos resultados, los cuales facilitan el proceso a medida que avanza. Estos experimentos son un total de 66, sumando un tiempo de experimentación de 33 horas.

5.4. Modelo final

Una vez concluida la fase de experimentación, se realiza el proceso de selección, de un modelo final, el cual debe ser el que mejor generaliza el problema y que en este caso clasifique mejor las emociones del conductor.

Dado que se trata de una clasificación múltiple, se utilizan la métrica de la precisión y la matriz de confusión obtenidos en el conjunto de test para seleccionar el mejor modelo. También hay que tener en cuenta que se utilizan conjuntos y diferente formato de los datos utilizados en el entrenamiento, los cuales son factores que hay que sopesar a la hora de decidir si ese modelo es el que mejores resultados aportaría en un contexto real.

En las TABLA 5.4 y TABLA 5.5 donde se muestran los resultados de la fase de experimentación, se puede observar como para cualquiera de los dos tipos de datos, el subconjunto de datos que mejores resultados aporta es aquel que no contiene instancias en alemán, lo cual puede deberse a las claras diferencias fonéticas entre ambos idiomas y la representación minoritaria de este conjunto en los conjuntos de datos.

Comparando las aproximaciones de los distintos formatos de datos en el subconjunto con instancias únicamente en inglés, la aproximación que utiliza imágenes ofrece unos resultados significativamente mejores sobre el mismo subconjunto de datos. En los conjuntos sin embargo que contiene instancias en alemán, la aproximación con matrices de valores decimales obtiene resultados de una media de un 10% mejores. A pesar de esta mejora, estos resultados son muchos inferiores comparados con los del dataset *RT_I*.

Debido a esto, el modelo seleccionado es el asociado al experimento 17, el cual se encuentra destacado en la TABLA 5.4. Este modelo tiene una precisión del **93.34%** en la tarea de clasificar 7 emociones. Su matriz de confusión se muestra en la Fig. 5.2, la cual muestra las clasificaciones realizadas por el modelo frente a las clases verdaderas. Los números que nombran las filas y columnas se corresponden con la codificación establecida en la Fig. 4.1 del apartado de extracción de datos y etiquetado. La diagonal principal marcada muestra signos de una buena clasificación para las siete emociones clasificadas.

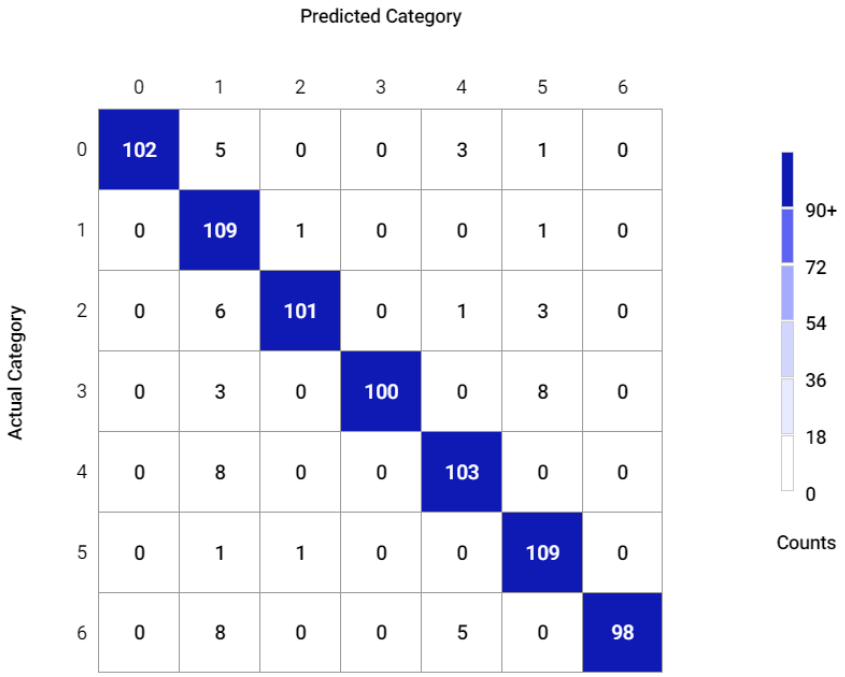


Fig. 5.2. Matriz de confusión del modelo final

A parte de la matriz de confusión también se muestran la gráfica de la evolución de la precisión en el entrenamiento en la Fig. 5.3 y la evolución del valor de la función de error en la Fig. 5.4. En ellas se puede apreciar que no hay signos de sobreentrenamiento y como el conjunto de test es casi igual de generalizado como el de entrenamiento.

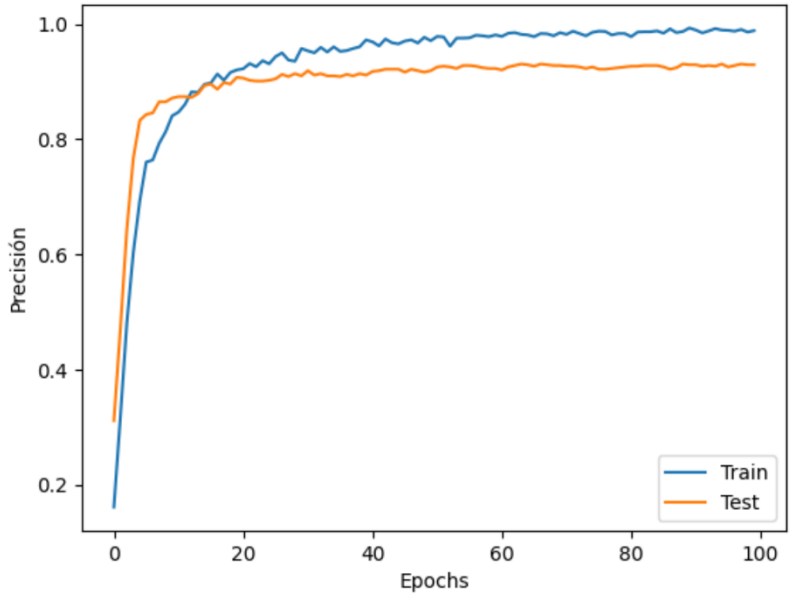


Fig. 5.3. Comparación entre la evolución de la precisión en los conjuntos de test y train durante el entrenamiento

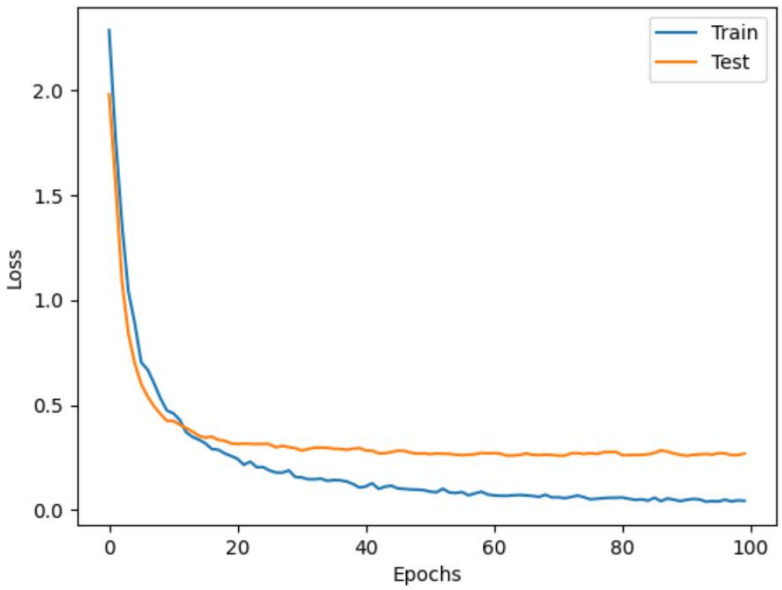


Fig. 5.4. Comparación de la evolución del error en los conjuntos de test y train durante el entrenamiento

Por último cabe remarcar que este modelo es escogido para utilizarlo en el contexto para el cual se entrena, es decir, cuando las conversaciones que se producen dentro de un coche son en inglés. Esta decisión se realiza debido a los malos resultados obtenidos en modelos en los cuales se utilizan más de un idioma y a la posibilidad de entrenar cada modelo en distintos idiomas dependiendo del contexto en el que se aplique.

5.5. Comparación entre aproximaciones

A pesar de resolverse mediante la misma técnica, las dos aproximaciones del formato de los datos presentan diferencias sustanciales tanto en los resultados como en la manera de obtenerlos.

La aproximación que se elige para la solución final es la que trata con imágenes, pero hay aplicaciones donde no sería la mejor opción. Utilizar imágenes para entrenar los modelos ha demostrado aportar información extra que mejora el desempeño de los modelos, lo que permite tener una precisión del 93%. Sin embargo, cuando se introducen audios en alemán son las aproximaciones en matrices de valores decimales las que aportan una precisión mayor.

Otra desventaja del uso de imágenes es el coste computacional que conlleva trabajar con ellas, ya que su representación gráfica añade gran cantidad de píxeles y además añade 2 canales más debido a su representación en RGB. Tras el procesado que convierte las matrices en imágenes el número de datos que analizan las redes convolucionales se multiplica por 380, generando un aprendizaje mucho más lento y la necesidad de hardware con mayor capacidad de almacenamiento. A parte del aprendizaje, también ralentiza la predicción ya que añade al preprocesado la conversión a imagen y el mayor número de datos tarda más en propagarse por la red y obtener una predicción.

En conclusión, el uso de imágenes aporta mejores resultados, pero requiere de más recursos de computación y memoria; mientras que el uso de matrices de valores decimales aporta mejores resultados para los subconjuntos de datos con dos idiomas, y puede ser aplicable para contextos en los que los recursos de hardware y el tiempo de respuesta sean limitados.

6. EVALUACIÓN DEL FUNCIONAMIENTO DEL SISTEMA

Una vez implementado el sistema, es necesario comprobar que funciona correctamente según los requisitos especificados en su diseño. Esta comprobación se realiza mediante la definición de un plan de pruebas y su realización en la práctica que permita asegurar el correcto funcionamiento del sistema evaluado.

6.1. Plan de pruebas

Para definir el plan de pruebas se utiliza un diseño tabular donde se especifica cada una de las pruebas realizadas utilizando la plantilla que se muestra en la TABLA 6.1.

TABLA 6.1. PLANTILLA DE PRUEBAS DEL SISTEMA

Identificador	**P-XX**
Descripción	
Objetivo	
Resultado esperado	
Resultado obtenido	
Requisitos asociados	

Cuyos atributos son los siguientes:

- **Identificador:** código único que identifica el sistema. El valor **XX** es un número de dos cifras que se va incrementando en una unidad para cada prueba.
- **Descripción:** breve especificación de cómo se lleva a cabo la prueba.
- **Objetivo:** breve descripción de qué función del sistema se quiere probar.
- **Resultado esperado:** salida que debe devolver el sistema en caso de funcionar correctamente.
- **Resultado obtenido:** salida que devuelve el sistema al realizar la prueba.
- **Requisitos asociados:** identificadores de los requisitos definidos en el diseño del sistema cuya funcionalidad es puesta a prueba.

Debido a la gran subjetividad asociada a las emociones y la dificultad de generar un conjunto para las pruebas en el entorno real, se ha utilizado el dataset *RBT_I* el cual contiene pistas generadas por actores con formación. Esto apenas afecta a la realización de las pruebas ya que la mayoría de ellas tienen como objetivo verificar el correcto funcionamiento del tratamiento de la entrada, lo cual es independiente de la emoción correspondiente.

Las pruebas y sus resultados se detallan a continuación utilizando la plantilla descrita anteriormente:

TABLA 6.2. PRUEBA DEL SISTEMA P-01

Identificador	**P-01**
Descripción	El sistema recibe un archivo *.wav* y lo carga correctamente
Objetivo	Comprobar la correcta recepción de archivos de audio
Resultado esperado	El sistema carga el archivo sin producir ningún error
Resultado obtenido	El sistema carga el archivo sin producir ningún error
Requisitos asociados	RF-01

TABLA 6.3. PRUEBA DEL SISTEMA P-02

Identificador	**P-02**
Descripción	El sistema recibe un audio previamente cargado con fragmentos en silencio
Objetivo	Comprobar la correcta eliminación de los fragmentos en silencio
Resultado esperado	Se eliminan las partes en silencio
Resultado obtenido	Se eliminan las partes en silencio
Requisitos asociados	RF-02

TABLA 6.4. PRUEBA DEL SISTEMA P-03

Identificador	**P-03**
Descripción	El sistema recibe fragmentos de audio con distintas frecuencias de muestreo
Objetivo	Comprobar el correcto procesado de señal con cualquier valor de frecuencia de muestreo
Resultado esperado	Se procesa correctamente el audio
Resultado obtenido	Se procesa correctamente el audio
Requisitos asociados	RF-03

TABLA 6.5. PRUEBA DEL SISTEMA P-04

Identificador	**P-04**
Descripción	El sistema recibe fragmentos de audio con distinto registros de voz, tanto agudos como graves y con distintos timbres
Objetivo	Comprobar el correcto procesado y predicción de señales con cualquier valor del registro de voz
Resultado esperado	Se procesa y predice correctamente
Resultado obtenido	Se procesa y predice correctamente
Requisitos asociados	RF-04

TABLA 6.6. PRUEBA DEL SISTEMA P-05

Identificador	**P-05**
Descripción	El sistema recibe fragmentos de audio en distintos idiomas
Objetivo	Comprobar el correcto procesado de señal con cualquier idioma
Resultado esperado	Se procesa y predice correctamente
Resultado obtenido	Se procesa correctamente y se predice correctamente el 55% de las veces
Requisitos asociados	RF-05

TABLA 6.7. PRUEBA DEL SISTEMA P-06

Identificador	**P-06**
Descripción	El sistema recibe fragmentos de audio de varias personas en una conversación
Objetivo	Comprobar el correcto procesado de señales de distintas fuentes
Resultado esperado	Se procesa y predice correctamente
Resultado obtenido	Se procesa y predice correctamente
Requisitos asociados	RF-06

TABLA 6.8. PRUEBA DEL SISTEMA P-07

Identificador	**P-07**
Descripción	El sistema recibe un fragmento de audio y se observa el producto del procesado
Objetivo	Comprobar el correcto procesado de señales en espectrogramas
Resultado esperado	Se procesa y se crea el espectrograma correspondiente correctamente
Resultado obtenido	Se procesa y se crea el espectrograma correspondiente correctamente
Requisitos asociados	RF-07

TABLA 6.9. PRUEBA DEL SISTEMA P-08

Identificador	**P-08**
Descripción	El sistema recibe un fragmento de audio y se observa la salida del sistema
Objetivo	Comprobar que la señal se está convirtiendo en una emoción
Resultado esperado	Se procesa y se devuelve la predicción de una emoción
Resultado obtenido	Se procesa y se devuelve la predicción de una emoción
Requisitos asociados	RF-08

TABLA 6.10. PRUEBA DEL SISTEMA P-09

Identificador	**P-09**
Descripción	Se realiza una predicción y se observa que se devuelve un valor de emoción
Objetivo	Comprobar que la señal se devuelve en forma de emoción para que el sistema actuador la utilice de entrada
Resultado esperado	Se devuelve la predicción de una emoción
Resultado obtenido	Se devuelve la predicción de una emoción
Requisitos asociados	RF-09

Una vez definidas las pruebas del sistema y con la intención de comprobar que todos los requisitos del sistema han sido verificados por las pruebas, se realiza una matriz de trazabilidad, la cual se muestra en la TABLA 6.11, donde se muestran las pruebas en las filas y los requisitos asociados en las columnas. En caso de que una prueba esté asociada al requisito se añade una X a esa casilla.

TABLA 6.11. MATRIZ DE TRAZABILIDAD ENTRE PRUEBAS Y REQUISITOS DEL SISTEMA

Requisitos / Pruebas	**RF-01**	**RF-02**	**RF-03**	**RF-04**	**RF-05**	**RF-06**	**RF-07**	**RF-08**	**RF-09**
P-01	**X**								
P-02		**X**							
P-03			**X**						
P-04				**X**					
P-05					**X**				
P-06						**X**			
P-07							**X**		
P-08								**X**	
P-09									**X**

El resultado obtenido de la matriz de trazabilidad entre requisitos y pruebas del sistema es satisfactorio ya que todos los requisitos están asociados por lo menos a una prueba del sistema, por lo que se concluye que el sistema de pruebas es completo y verifica todas las funcionalidades que ofrece el sistema.

7. GESTIÓN DEL PROYECTO

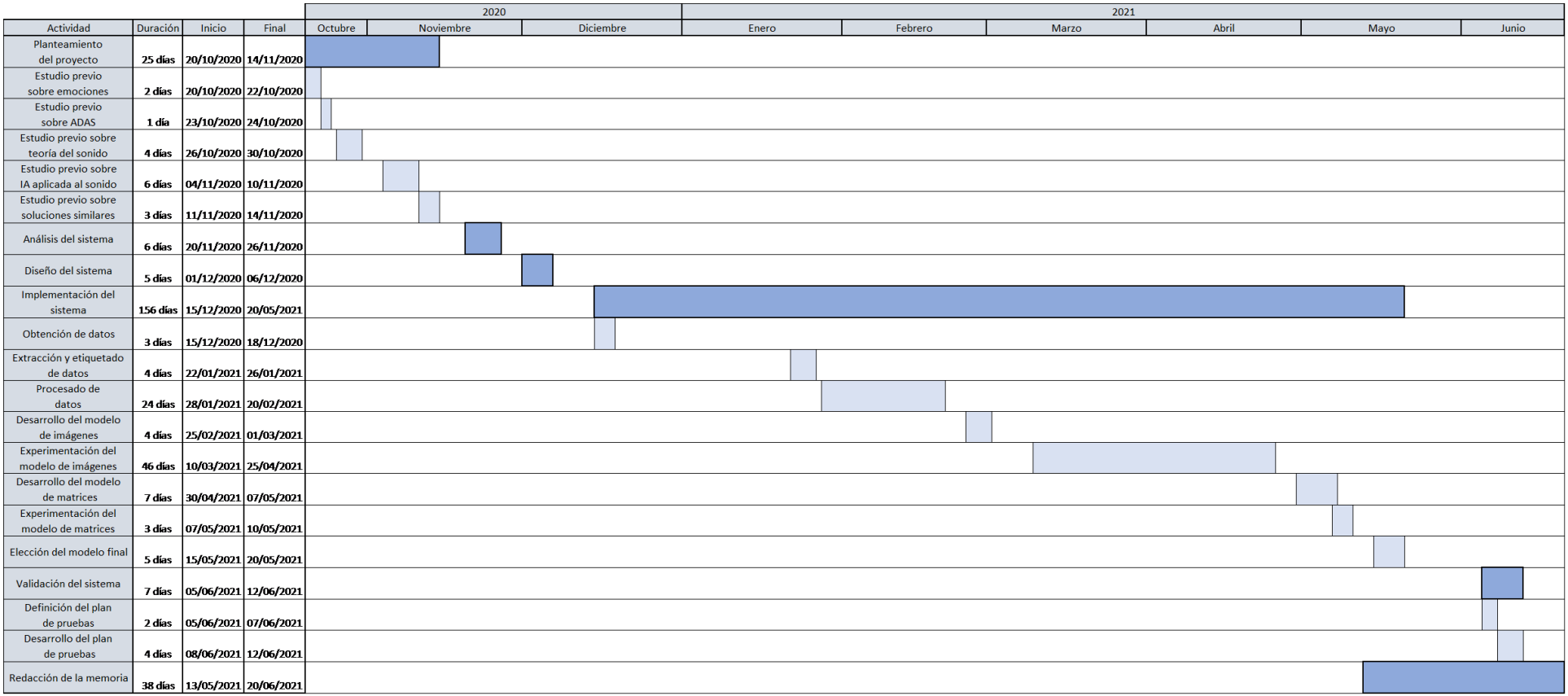
7.1. Planificación

En este apartado se muestran las actividades realizadas a lo largo del proyecto, su fecha de inicio, de final y el número de horas dedicadas a esa actividad. La TABLA 7.1 muestra esta información, donde las filas sombreadas son las actividades principales de un proyecto de minería de datos y de la realización de este proyecto; y las que están sin sombrear son las subactividades de cada una de esas actividades principales.

TABLA 7.1. LISTADO DE ACTIVIDADES REALIZADAS

Actividad	Fecha de inicio	Fecha de finalización	Horas dedicadas
Planteamiento del proyecto	**20/10/2020**	**14/11/2020**	**65**
Estudio previo sobre emociones	20/10/2020	22/10/2020	10
Estudio previo sobre ADAS	23/10/2020	24/10/2020	5
Estudio previo sobre teoría del sonido	26/10/2020	30/10/2020	15
Estudio previo sobre IA aplicada al sonido	4/11/2020	10/11/2020	20
Estudio previo sobre soluciones similares	11/11/2020	14/11/2020	15
Análisis del sistema	**20/11/2020**	**26/11/2020**	**10**
Diseño del sistema	**1/12/2020**	**6/12/2020**	**8**
Implementación del sistema	**15/12/2020**	**20/5/2021**	**215**
Obtención de datos	15/12/2020	18/12/2020	5
Extracción y etiquetado de datos	22/1/2021	26/1/2021	10
Procesado de datos	28/1/2021	20/2/2021	80
Desarrollo del modelo de imágenes	25/2/2021	1/3/2021	15
Experimentación del modelo de imágenes	10/3/2021	25/4/2021	40
Desarrollo del modelo de matrices	30/4/2021	7/5/2021	15
Experimentación del modelo de matrices	7/5/2021	10/5/2021	10
Elección del modelo final	15/5/2021	20/5/2021	10
Validación del sistema	**5/6/2021**	**12/06/2021**	**30**
Definición del plan de pruebas	5/6/2021	7/6/2021	10
Desarrollo del plan de pruebas	8/6/2021	12/6/2021	20
Redacción de la memoria	**13/05/2021**	**20/6/2021**	**65**

A partir de esta tabla se deriva un diagrama de Gantt, el cual permite exponer de una manera más visual el tiempo dedicado a cada actividad realizada. El diagrama de Gantt se encuentra en la Fig. 7.1 y muestra la planificación con una granularidad de meses. En él se puede apreciar como de los ocho meses y medio en los que se extiende el proyecto, la implementación del sistema toma la mayor fracción de tiempo. En ella las actividades que más tiempo llevan son el procesado de datos y la experimentación del modelo de imágenes, ya que son las partes donde más información y pruebas hay que realizar para obtener buenos resultados.



Actividad	Duración	Inicio	Final
Planteamiento del proyecto	25 días	20/10/2020	14/11/2020
Estudio previo sobre emociones	2 días	20/10/2020	22/10/2020
Estudio previo sobre ADAS	1 día	23/10/2020	24/10/2020
Estudio previo sobre teoría del sonido	4 días	26/10/2020	30/10/2020
Estudio previo sobre IA aplicada al sonido	6 días	04/11/2020	10/11/2020
Estudio previo sobre soluciones similares	3 días	11/11/2020	14/11/2020
Análisis del sistema	6 días	20/11/2020	26/11/2020
Diseño del sistema	5 días	01/12/2020	06/12/2020
Implementación del sistema	156 días	15/12/2020	20/05/2021
Obtención de datos	3 días	15/12/2020	18/12/2020
Extracción y etiquetado de datos	4 días	22/01/2021	26/01/2021
Procesado de datos	24 días	28/01/2021	20/02/2021
Desarrollo del modelo de imágenes	4 días	25/02/2021	01/03/2021
Experimentación del modelo de imágenes	46 días	10/03/2021	25/04/2021
Desarrollo del modelo de matrices	7 días	30/04/2021	07/05/2021
Experimentación del modelo de matrices	3 días	07/05/2021	10/05/2021
Elección del modelo final	5 días	15/05/2021	20/05/2021
Validación del sistema	7 días	05/06/2021	12/06/2021
Definición del plan de pruebas	2 días	05/06/2021	07/06/2021
Desarrollo del plan de pruebas	4 días	08/06/2021	12/06/2021
Redacción de la memoria	38 días	13/05/2021	20/06/2021

Fig. 7.1. Diagrama de Gantt del proyecto

7.2. Presupuesto

En este apartado se calculan y muestran los costes asociados a la ideación y desarrollo de este proyecto. La tarea de realizar el presupuesto es de fundamental importancia, ya que en un mercado donde existe gran competitividad, es crucial realizar una buena estimación de costes y crear una oferta para que posibles clientes la acepten y se generen beneficios económicos.

El cálculo del presupuesto se encuentra desglosado en los diferentes tipos de coste que se pueden encontrar en el proyecto: de personal, de hardware, de software y costes indirectos.

Los costes de personal van asociados a los recursos humanos utilizados. Al tratarse de un proyecto de ingeniería complejo, es necesaria la creación de un equipo de trabajo que permita dividir las responsabilidades entre personas con cualificaciones específicas de ciertas tareas. En concreto se requieren de dos perfiles distintos: jefe de proyecto, el cual se encarga de planificar, supervisar y gestionar el proyecto desde una perspectiva general; y data scientist, cuya responsabilidad es analizar el problema, idear una solución basada en minería de datos e implementarla.

Los costes de personal asociados a estos puestos se muestran en la TABLA. 7.2, cuya información ha sido obtenida del portal online de oferta de empleo Tecnoempleo [44], el cual calcula el salario medio a través de las propias ofertas que se realizan en su plataforma y del cual se asume que pertenece a una jornada completa de 8 horas diarias. Dado que los sueldos aportados por esta fuente son anuales, se realiza una estimación asociada a la duración de este proyecto con la siguiente fórmula:

$$Coste_{empleado}(€) = \frac{Salario_{empleado}(€/año)}{12\ (meses/año) \cdot\ 20(días/mes) \cdot 8(horas/día)} \cdot horas_{empleado}$$

TABLA 7.2. COSTES DE PERSONAL DEL PROYECTO

Puesto	N.º empleados	Horas	Salario anual	Coste
Jefe de proyecto	1	128	37.800 €	**2.520 €**
Data scientist	1	265	35.000 €	**4.830 €**
Total				**7.350 €**

Los costes de hardware que se muestran en la TABLA 7.3 hacen referencia a los equipos físicos utilizados para el desarrollo de todos los productos asociados al proyecto. Dado que los equipos no se han comprado específicamente para el desarrollo de este proyecto y seguirán siendo utilizados tras la finalización de este, es necesario calcular el precio asociado a la depreciación del equipo durante la realización del proyecto. Para

calcular este valor es necesario saber la vida útil del equipo y utilizar la fórmula que se muestra a continuación:

$$Coste_{equipo}(€) = \frac{precio_{equipo}(€)}{vida\ útil(mes)} \cdot utilización(mes)$$

TABLA 7.3. COSTES DE HARDWARE DEL PROYECTO

Equipo	Precio	Vida útil	Utilización	Coste
Ordenador de desarrollo	930 €	72 meses	9 meses	**116,25 €**
Ordenador de desarrollo portátil	1.700 €	60 meses	4 meses	**113,3 €**
Ordenador de experimentación	4.200 €	84 meses	4 meses	**200 €**
Total				**429,55 €**

Los costes de software que se muestran en la TABLA 7.4 son aquellos asociados a las licencias de los productos de software. Estas licencias pueden tener subscripciones que requieren renovarlas cada cierto tiempo, ser una licencia de por vida o ser software libre el cual no requiere ningún gasto.

TABLA 7.4. COSTES DE SOFTWARE DEL PROYECTO

Software	Tipo de licencia	Precio	Utilización	Coste
Pycharm	Mensual	20 €/mes	6 meses	**120 €**
Windows 10 Pro	Definitiva	11,9 €	9 meses	**11,9 €**
Linux 5.8.0	Libre	0 €	4 meses	**0 €**
Microsoft office	Definitiva	20 €	9 meses	**20 €**
CometML	Libre	0 €	4 meses	**0 €**
Keras/Tensorflow	Libre	0 €	4 meses	**0 €**
Librosa	Libre	0 €	6 meses	**0 €**
Multiprocessing	Libre	0 €	6 meses	**0 €**
Numpy	Libre	0 €	6 meses	**0 €**
Total				**151,9 €**

El último tipo de coste son los gastos indirectos, los cuales no se pueden imputar a un producto en concreto, como son la luz y el agua o la conexión a internet. Estos se encuentran especificados en la TABLA 7.5.

TABLA 7.5. COSTES INDIRECTOS DEL PROYECTO

Nombre	Precio	Duración	Coste
Luz y agua	37 €/mes	9 meses	**333 €**
Conexión a internet	18 €/mes	9 meses	**162 €**
Total			**495 €**

La suma de todos estos costes se utiliza para calcular el coste total del proyecto, para el cual hay que tener en cuenta un margen de riesgo para posibles imprevistos y un margen

de beneficio. A parte de estos porcentajes extra, también hay que tener en cuenta el porcentaje de IVA correspondiente. Este cálculo y coste total se muestra en la TABLA 7.6.

TABLA 7.6. COSTE TOTAL DEL PROYECTO

Coste total del proyecto	
Costes directos	7.930,55 €
Costes indirectos	495 €
Margen de riesgo (15%)	1.189,58 €
Margen de beneficio (20%)	1.586,11 €
Coste sin IVA	11.201,24 €
IVA 21%	2.352,26 €
Coste con IVA	**13.553,5 €**

7.3. Impacto socio-económico

Las nuevas tecnologías que se desarrollan día a día tienen impactos directos en la sociedad que pueden ser positivos o negativos, los cuales a su vez influyen en la industria que las crea. Por este motivo es importante estudiar cuál va a ser el impacto que va a tener el sistema desarrollado en lo que concierne a la sociedad y a los propios intereses económicos de la industria, ya que va a influir en su futuro en cierta medida.

Los sistemas de ayuda a la conducción tienen una peculiaridad que comparten industrias como la farmacéutica o sanitaria, que es el objetivo de reducir o mitigar las fatalidades que acaban con vidas humanas. La Dirección General de Tráfico (DGT) estima que la media del coste económico anual de accidentes de tránsito es del 1% del PIB en países de ingresos bajos y hasta del 2% en países con ingresos altos [45], aparte de los costes psicológicos y en vidas humanas.

Los sistemas de ayuda a la conducción reducen ese coste tan elevado, beneficiando económicamente a toda la sociedad que los utiliza, ya que un accidente de tráfico conlleva costes que se pagan mediante impuestos como son los servicios sanitarios, las reparaciones de los daños causados en la red de carreteras nacionales o los servicios de emergencias, en el caso de España.

Otro factor social afectado por la reducción de los accidentes de tráfico que podría conseguir la inclusión de este sistema en la práctica es el medioambiental. Este factor se ve afectado dado que muchos de los accidentes que se producen se dan en zonas donde la carretera atraviesa un bosque o una zona de interés medioambiental, los cuales al producirse provocan vertidos de sustancias nocivas o incendios que afectan negativamente al ecosistema donde suceden.

A parte de los impactos que tiene en la sociedad este tipo de sistema, también es necesario estudiar su impacto económico en beneficio de la compañía u organización que lo desarrolla. Este factor es importante ya que en un sector tan competitivo como es el

tecnológico, crear un producto que aporte una mejora del entorno de la empresa o una mejora para sus propios productos a parte de beneficios económicos directos, aumenta los beneficios y posibilidades futuras.

Uno de estos beneficios económicos indirectos es la creación de un nuevo ADAS que requiere de puestos de trabajo para su desarrollo, pero también para su homologación y supervisión.

Otro beneficio indirecto viene a raíz de la posibilidad de utilizar la emoción detectada del conductor para adaptar el funcionamiento de otros ADAS para mejorar su efectividad. Esta aplicación crearía un beneficio en la calidad de otros sistemas y crearía los puestos de trabajo necesarios para adaptarlos.

En resumen, la introducción de este sistema en la práctica provocaría efectos sociales positivos mediante la reducción de muertes y lesiones provocadas por los accidentes de tráfico, y también efectos económicos positivos al crear nuevos posibles productos a los cuales los desarrolladores del sistema podrían obtener beneficio.

8. CONCLUSIONES

8.1. Conclusiones generales

Tras realizar la totalidad del proyecto y la redacción de la memoria correspondiente, se puede concluir que el objetivo principal y los objetivos específicos se han conseguido satisfactoriamente.

El estudio del Estado del Arte ha aportado una visión general de la situación actual en el ámbito de los ADAS, el uso de técnicas de procesado de audio y el uso de técnicas de *Deep Learning* para la clasificación de emociones. También se ha obtenido una base para la realización del proyecto a partir de la información aportada por trabajos similares.

El proceso de diseño y experimentación ha concluido con el desarrollo de un sistema funcional de procesamiento, el cual podría formar parte de un ADAS que utilizara sus predicciones para tomar decisiones o adaptar su propio funcionamiento.

La fase de experimentación ha producido un total de 58 experimentos válidos, con un tiempo de entrenamiento cercano a las 16 horas. Este proceso ha permitido también comparar dos aproximaciones distintas con respecto al formato de la entrada de los modelos, diferenciando entre imágenes y matrices de valores decimales.

En la fase de evaluación del sistema, se crean y se muestran los resultados de una serie de pruebas que verifican el correcto funcionamiento del sistema según las especificaciones definidas en los requisitos.

El desarrollo de este proyecto ha servido para aprender competencias relevantes de cara a abordar un caso real que pueda generar beneficios sociales y económicos. Estas competencias adquiridas son la planificación y realización completa de un proyecto de minería de datos, la comprensión del contexto de un problema, la adquisición y procesado de un conjunto de datos y la creación de los modelos que generalicen las instancias que se den en un contexto real.

En último lugar, este proyecto a parte de servir para completar el proceso formativo del alumno puede aportar conceptos nuevos al estado del arte asociado, sobre todo en la comparación entre distintas metodologías y aproximaciones alternativas en el tratado de datos de sonido.

8.2. Limitaciones y dificultades encontradas

Los trabajos que conllevan la utilización de las emociones humanas contienen una serie de problemas que se asocian a la dificultad de definirlas, identificarlas y estudiarlas. Esta complejidad proviene en gran parte de la naturaleza inexacta de las emociones humanas que se ven afectadas por gran cantidad de factores como pueden ser la personalidad, el contexto, el estado físico, etc.; lo cual complica una tarea de reconocimiento que incluso las personas muchas veces consideran complicada. Si a la

dificultad de reconocer las emociones se le añade la restricción de utilizar solamente el habla, hace de ella una tarea de gran complejidad.

Por otra parte, los conjuntos de datos que se utilicen afectan muy significativamente y al tratarse de datos con muchos atributos a tener en cuenta, generalizar un sistema para cualquier idioma, edad o situación constituye un reto. Además, la obtención de bases de datos etiquetadas es complejo ya que escasean y las que son accesibles pueden introducir sesgos debido a la gran variedad de maneras de crear un dataset de estas características y de etiquetar según unas clases que son en cierta medida subjetivas.

En resumen, la mayor dificultad ha sido generalizar un problema con tantos posibles formatos y características de datos, a pesar de que solamente se ha tenido en cuenta una pequeña porción de las posibles posibilidades.

8.3. Trabajos futuros

El producto final de este proyecto es la componente de procesado de un sistema de ayuda a la conducción, la cual necesitaría de la componente relacionada a la recogida del audio con un micrófono u otro tipo de dispositivo y también el sistema actuador que permitiera tomar decisiones a partir de la emoción del conductor y ponerlas en práctica. Por ello un trabajo futuro podría consistir en el desarrollo de este sistema al completo y su despliegue en un contexto real para evaluar su funcionamiento.

Otra ampliación de este trabajo podría ser estudiar cómo adaptar sistemas de ayuda a la conducción ya existentes dependiendo de la emoción obtenida. Esta adición es de gran interés ya que podría mejorar sistemas que ya se utilizan en una gran cantidad de vehículos, y que supondría una gran mejoría en la seguridad de los conductores y pasajeros.

Finalmente, este trabajo podría ser continuado a través del estudio de la influencia de los distintos idiomas en la detección de las emociones, ya que en este trabajo solamente se han utilizado dos idiomas y es posible que, encontrando cierta relación entre emoción en el habla e idioma, u otras características demográficas como sexo o edad, se consiguiera un sistema más preciso y con mayor capacidad de generalizar.

9. BIBLIOGRAFÍA

[1] M. Ruiza, T. Fernández y E. Tamaro. "Biografía de Karl Benz". Biografías y Vidas: La enciclopedia biográfica en línea. https://www.biografiasyvidas.com/biografia/b/benz.htm (acceso: 8 de mayo de 2021).

[2] M. Ruiza, T. Fernández y E. Tamaro. "Henry Ford. Biografía". Biografías y Vidas: La enciclopedia biográfica en línea. https://www.biografiasyvidas.com/monografia/ford/ (acceso: 8 de mayo de 2021).

[3] Eurostat. "Passenger cars in the EU". Eurostat: statistics explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Passenger_cars_in_the_EU#Overview (acceso: 8 de mayo de 2021)

[4] ACEA. "Report: Vehicles in use, Europe – January 2021". ACEA. https://www.acea.be/publications/article/report-vehicles-in-use-europe-january-2021. (acceso: 8 de mayo 2021)

[5] Organización Mundial de la Salud. "10 datos sobre la seguridad vial en el mundo". Organización Mundial de la Salud. https://www.who.int/features/factfiles/roadsafety/es/ (acceso: 8 de mayo 2021)

[6] A. Díaz. "Número de muertes por accidentes de tráfico en España de 2006 a 2019". Statista. https://es.statista.com/estadisticas/592222/numero-de-muertes-por-accidentes-de-trafico-en-espana/ (acceso: 8 de mayo 2021)

[7] L. Montoro. "El factor humano en la causalidad de los accidentes de Tráfico" presentada en el Congreso de los Diputados, Madrid, 9 jun., 2017. [En línea]. Disponible en: http://www.pat-apat.org/congreso-de-los-diputados-jornada-factor-humano-causalidad-accidentes-de-trafico/

[8] C. De Locht y H. Van Den Broeck. "Complementary metal-oxide-semiconductor (CMOS) image sensors for automotive applications". Science Direct. https://www.sciencedirect.com/topics/engineering/advanced-driver-assistance-systems#:~:text=ADAS%20systems%20include%20long%2D%20and,range%20radar%20and%20vision%20systems.&text=Active%20safety%20systems%20include%20adaptive,automatic%20steering%20and%20braking%20intervention (acceso: 9 de mayo 2021)

[9] M. Ratcliffe, "William James on Emotion and Intentionality", *International Journal of Philosophical Studies*, vol. 13, n. º 2, pp. 179-202, feb. 2005. [En línea]. Disponible en: https://d1wqtxts1xzle7.cloudfront.net/2035130/jamesproof.pdf?response-

content-disposition=inline%3B+filename%3DWilliam_James_on_Emotion_and_Intentional.pdf&Expires=1623834268&Signature=TeJ4FccKCiUFzxYIxzDn0tPUGplhatD8SKt50JfJHHWDJrVqehPoWoTL917CkniCFyqvgFBkx9Lb2HFiUu-L5ND6rnXTxE2VLBfOqRWkl~iREmOW2GJRUTyjub2ueJGJ3c0SbDZSmyKrR1X4M2~6j4y-bF3bAgQ83FQ78HvxykuPBBjGvUF-OZnJCvF0w4kNdcc0yuZuxqZDiazJNITdQuwPHTSqx5zhGyBq9RFrWp~o6z145STQjNU1LhUOe6GCONLwewv40w7Ws5wd-2vPEFMMOhbjgRMwIMTv8hFt5t1fcl66kVfPx49XxzyL8eSoUazYv8tzI-z36LHR~Cu08Q__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA Acceso: junio 2021

[10] W. B. Cannon, “The James-Lange Theory of Emotions: A Critical Examination and Alternative Theory”, *The American Journal of Psychology*, vol. 100, n. º 3-4, pp. 567-586, 1987. [En línea]. Disponible en: https://www.jstor.org/stable/1422695 Acceso: junio 2021

[11] D. G. Myers. *Psychology,* 6th ed. Nueva York: Worth Publishers, 2000

[12] V. Nitsch y M. Popp, “Emotions in robot psychology”, *Biological Cybernetics*, vol. 108, pp. 621-629, 2014. [En línea]. Disponible en: https://www.semanticscholar.org/paper/Emotions-in-robot-psychology-Nitsch-Popp/316a069eff3f8ac26f298c348318b25f17c67064 Acceso: junio 2021

[13] Y-S. Seo y J-H. Huh. “Automatic Emotion-Based Music Classification for Supporting Intelligent IoT Applications”. ResearchGate. https://www.researchgate.net/figure/Russells-circumplex-model-The-circumplex-model-is-developed-by-James-Russell-In-the_fig1_330817411#:~:text=The%20circumplex%20model%20is%20developed%20by%20James%20Russell.,to%20the%20intensity%20of%20emotion. (acceso: 11 de mayo)

[14] P. N. Juslin y K. R. Scherer, “Speech emotion analysis”, *Scholarpedia,* vol. 3, n.º 10, pp. 4240, oct. 2008. [En línea]. Disponible en: http://www.scholarpedia.org/article/Speech_emotion_analysis acceso: mayo 2021

[15] K. R. Scherer, “Vocal communication of emotion: A review of research paradigms”, *ScienceDirect,* vol. 40, n.º 1-2, pp. 227-256, abr. 2003. [En línea]. Disponible en: https://www.sciencedirect.com/science/article/pii/S0167639302000845#BIB123 Acceso: mayo 2021

[16] C. M. Tyng, H. U. Amin, M. N. M. Saad y A. S. Malik, “The influences of Emotion on Learning and Memory”, *Frontiers in Psychology,* vol. 8, pp. 1454,

ago. 2017. [En línea]. Disponible en: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01454/full Acceso: junio 2021

[17] K. Steinhauser et al., “Effects of emotion on driving behavior”, *ScienceDirect,* vol. 59, pp. 150-163, nov. 2018. [En línea]. Disponible en: https://www.sciencedirect.com/science/article/pii/S136984781830278X?casa_token=jzu3EY763LoAAAAA:vfiUcSAbSyHpTV-_aYUgzIQr9haWaZWHlUB08Pz1PsYVpI2WDgRtFSfz8BqoPtsz-RBBwzs7#bi005 Acceso: junio 2021

[18] SAE International, “Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles”, J3016_201806, junio, 15, 2018.

[19] km77. “Conducción autónoma | Los cinco niveles”. km77 https://www.km77.com/reportajes/varios/conduccion-autonoma-niveles (acceso: 2 de junio de 2021).

[20] OXTS. “ADAS Sensors”. OXTS. https://www.oxts.com/adas-sensors/ (acceso: 3 de junio de 2021).

[21] SAMSARA. “Advanced Driver Assistance Systems (ADAS) for Commercial Fleets”. SAMSARA. https://www.samsara.com/guides/adas (acceso: 3 de junio de 2021).

[22] M. Martínez-Ripoll. “Cristalografía Parte 5: Dispersión y difracción”. CSIChttps://www.xtal.iqfr.csic.es/Cristalografia/parte_05.html (acceso: 3 de junio de 2021)

[23] “Tema 1. Aspectos físicos del sonido”, apuntes de clase para Psicología de la Percepción, Departamento de Psicología, Universidad Complutense de Madrid, sin fecha.

[24] Tutorials Point. “Digitalization of sound”. Tutorials Point. https://www.tutorialspoint.com/multimedia/multimedia_sound_audio.htm (acceso: 3 de junio de 2021)

[25] IBM Cloud Education. “What is Artificial Intelligence (AI)?”. IBM. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence (acceso: 3 de junio de 2021)

[26] J. McCarthy. “What is Artificial Intelligence?”. Università Degli Studi di Milano. https://homes.di.unimi.it/borghese/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems_2008_2009/Old/IntelligentSystems_2005_2006/Documents/Symbolic/04_McCarthy_whatisai.pdf (acceso: 3 de junio de 2021)

[27] Oxford Lexico. “Inteligencia”. Lexico. https://www.lexico.com/es/definicion/inteligencia (acceso: 3 de junio de 2021)

[28] F. Correa, *AI Knowledge Map: How to Classify AI Technologies*, 1ª ed. Nueva York: Springer International Publishing, 2019. [En línea]. Disponible en: https://link.springer.com/chapter/10.1007/978-3-030-04468-8_4#citeas

[29] N. Bernache. “Artificial Intelligence, Machine Learning, and Deep Learning: Same context, Different concepts”. Master IESC Angers. https://master-iesc-angers.com/artificial-intelligence-machine-learning-and-deep-learning-same-context-different-concepts/ (acceso: 18 de junio de 2021)

[30] Significados. “Significado de neurona”. Significados https://www.significados.com/neurona/ (acceso: 4 de junio de 2021)

[31] P. Isasi. 2021. *Introducción a las redes de neuronas* [Presentación de PowerPoint]. Disponible en: https://aulaglobal.uc3m.es/.

[32] R. Mendoza. “Aplicación del gradiente en Redes Neuronales”. Medium https://medium.com/@ricardojmv85/aplicaci%C3%B3n-del-gradiente-en-redes-neuronales-78bff0d802d5 (acceso: 4 de junio de 2021)

[33] J. Sullivan. “Neural Network from Scratch: Perceptron Linear Classifier”. GitHub. https://jtsulliv.github.io/perceptron/ (acceso: 4 de junio de 2021)

[34] D. Rumelhart, G. Hinton y R. Williams, “Learning representations by back-propagating errors”, *Nature*, vol. 323, 533–536, may. 1986. [En línea]. Disponible en: https://doi.org/10.1038/323533a0 Acceso: junio 2021

[35] M. Wadhwa, A. Gupta y P. K. Pandey. “SPEECH EMOTION RECOGNITION (SER) THROUGH MACHINE LEARNING”. Analytics Insight. https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/ (acceso: 5 de junio de 2021)

[36] C. Jones y I. M. Jonsson. “Performance Analysis of Acoustic Emotion Recognition for In-Car Conversational Interfaces”. ResearchGate. https://www.researchgate.net/publication/339295916_Performance_Analysis_of_Acoustic_Emotion_Recognition_for_In-Car_Conversational_Interfaces (acceso: 7 de junio de 2021)

[37] K. Dupuis y M. K. Pichora-Fuller, 2010, “Toronto emotional speech set (TESS)” University of Toronto TSpace, doi: https://doi.org/10.5683/SP2/E8H2MF.

[38] S. R. Livingstone y F. A. Russo, 2018, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of

facial and vocal expressions in North American English" PLoS ONE 13, doi: https://doi.org/10.1371/journal.pone.0196391.

[39] W. Sendlmeier, F. Burkhardt, M. Kienast, A. Paeschke y B. Weiss, 1999, "Berlin Database of Emotional Speech" Technical University of Berlin, doi: https://www.kaggle.com/piyushagni5/berlin-database-of-emotional-speech-emodb.

[40] D. Martinez-Zorrila. "Los sintetizadores: una breve introducción". ResearchGate. https://www.researchgate.net/figure/Grafico-de-la-onda-senoidal_fig4_260403799 (acceso: 7 de junio de 2021)

[41] P. Orland. "Concept Fourier series in category python". Manning. https://livebook.manning.com/concept/python/fourier-series (acceso: 8 de junio de 2021)

[42] Wikipedia. "Escala Mel". Wikipedia. https://es.wikipedia.org/wiki/Escala_Mel (acceso: 9 de junio de 2021)

[43] C. E. Nwankpa, W. Ijomah, A. Gachagan y S. Marshall. "Activation Functions: Comparison of Trends in Practice and Research for Deep Learning. Arxiv.org. https://arxiv.org/pdf/1811.03378.pdf (acceso: 11 de junio de 2021).

[44] Tecnoempleo. "Informe Empleo Informática - junio 2021". Tecnoempleo. https://www.tecnoempleo.com/informe-empleo-informatica.php (acceso: 13 de junio de 2021)

[45] B. Zúñiga Reinares. "Tema 5: LAS CONSECUENCIAS DEL ACCIDENTE. CONSECUENCIAS INDIVIDUALES Y SOCIALES. COSTOS ECONÓMICOS. LAS VÍCTIMAS: CONCEPTOS Y TIPOLOGÍAS. AMBITOS DE ACTUACIÓN SOBRE VÍCTIMAS". DGT. https://www.dgt.es/Galerias/la-dgt/empleo-publico/oposiciones/doc/2014/TEMA-1.5.doc (acceso: 13 de junio de 2021)

10. ANEXO A: EXTENDED ABSTRACT

INTRODUCTION

Since its invention in 1886 by Karl Friedrich Benz [1] and its later mass production by Henry Ford [2], the gas car has changed society and economy globally. In 2018, almost 50% of Europe's population owned a car and this ratio is predicted to increase every year [3]. The introduction of cars in everyday life has increased society's quality, however it also has brought countless deaths and injuries, being the main cause of death in population between 15 and 29 years old [5].

In order to reduce these accidents which are mostly produced by human errors, most of the new security measures are introduced through Advanced Driving Assistance Systems (ADAS), which function is helping the driver to reduce the possibilities of having an accident caused by a distraction by introducing some level of automatization. These ADAS are mostly focused on the driver of the vehicle as is the one who is in charge of controlling it, but the driver's state is also changed by the influence of the other passengers. Because of this, a solution based on Artificial Intelligence techniques is proposed to monitor the driver's and passengers' emotion so other systems can take the necessary measures to avoid an accident.

OBJECTIVES

Considering traffic accidents as one of the principal causes of non-natural deaths and its socio-economic impact, the main objective is to create a component that could be implemented in an ADAS that recognizes the emotion of the people inside a car using Affective Computing and Deep Learning techniques. In order to achieve this, it is also necessary to accomplish a series of specific objectives such as studying the effect of emotions in driving, study and apply audio feature extraction techniques, define a model capable of recognizing emotions from speech and compare different data approaches.

STATE OF ART

Emotions and speech

An emotion can be defined by a state which predispose how a person adapts to the environment and the situations in it, which triggers a set of visible changes called sentiments. There are several theories that aim to classify emotions but there are two main ones: the discrete classification one given by Paul Eckman [12] which defines a uni-dimensional classification of basic emotions; and the circumplex theory given by James A. Russell, which classifies emotions in a two-axis scale where valence and arousal define each emotion.

One of the noticeable changes that emotion affects in humans is the speech. The physiological factors that are involved in speech are heavily affected by the emotional

state of the speaker, which means that speech is influenced by emotions. These factors are crucial in order to identify the emotion associated with a speaker. A relevant study related with this is the one made by Klaus R. Scherer and Tom Johnstone, whose observation of the speech characteristics of people that were induced with a certain emotion, concluded that there is a relation between specific physiological factors and specific emotions.

Influence of emotions in driving

Attention is a fundamental factor in driving, because a distraction at such high speed makes the reaction distance bigger and sometimes uncontrollable, which ends up being a car accident. Although attention is not considered to be an emotion, it is heavily affected by them and consequently affects someone's driving [16].

A study developed by Steinhauser et al. [17] researched about the positive and negative influence of emotions in drivers that were performing different activities related to driving such as braking or accelerating. Some of the relevant results from this study were that it is possible to influence emotion through imagination or music, there are some emotions which influence in driving activities is increased depending on the attention of the driver and that the effect of emotions is different depending on the activity.

Sound theory

Sound is a series of waves that are transmitted through an elastic medium, which in the case of speech is the air. These waves are characterized by a series of factors such as its amplitude, which is the distance from the highest part of the wave to the lowest, wavelength, which is the distance from same spots of two consecutive waves, or frequency, which is the number of cycles done by the wave in one second. These parts are shown in Fig. 10.1.

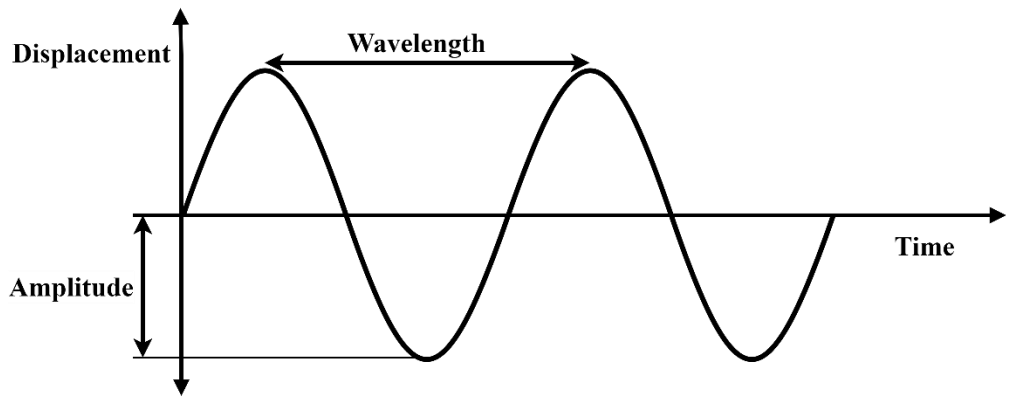


Fig. 10.1. Parts of a wave

Also, they have other characteristics associated with human sound perception such as intensity, pitch or tone, which allow humans to identify different sources of sound.

Sound digitalization

In order to use sound in an informatic system, it needs to be processed in a way that it is understandable by a computer. The beginning of this process is performed by a microphone, which measures the air pressure made by sound waves and transforms it into an analog signal. This continuous signal is then sampled by taking the value of the amplitude at specific periodic time instants. The number of samples taken is determined by the sampling rate. Once the wave is sampled, the value of the measured voltage is transformed into an interval of real numbers that represent the value of the amplitude in order to store them in 8 to 16 bits each. Once this process called quantification is finished, these values are encoded to binary values which can be processed by the computer. This whole process is shown in Fig. 10.2.

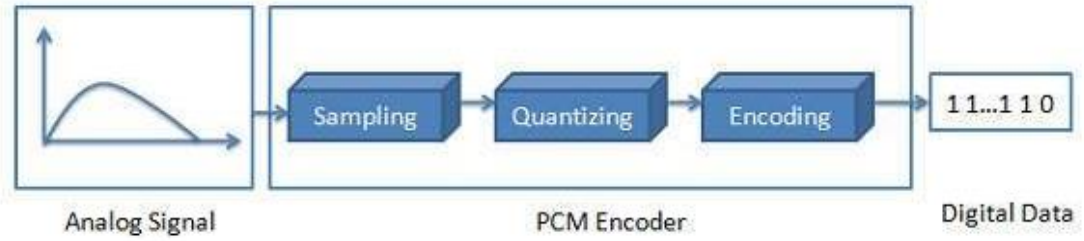


Fig. 10.2. Sound digitalization process [24]

Speech emotion recognition through Deep Learning

One of the multiple applications that Deep Learning has is speech emotion recognition because of its capability of extracting knowledge of big data inputs using several layers in their architectures. Convolutional Neural Networks are a specific technique inside Deep Learning that using convolutional filters combined with a densely connected Multilayer perceptron, can work with data with massive numbers of features and learn how to extract information from them.

One example of application is the one given by M. Wadhwa, A. Gupta y P. K. Pandey [35] in their work about using CNNs to classify emotions using speech. This project shows the complexity of recognizing emotions because of its variability through different people and the difficulty of finding universal features able to differentiate emotions. The data to train the CNN is extracted from five different sources and processed in order to extract features that help in the learning process. Its final result is a model that recognizes the emotion with a 75% accuracy for test data.

Another related work is the one made by C. M. Jones y I-M. Jonsson [36] that was motivated by the rise of conversational in-car assistances that could be improved through emotion recognition. The data used to train the models were generated from recording driving sessions and label them through questionnaires. After training the models, they achieved a 65% accuracy but with 15% of the wrongly predicted emotions being equivalent with the real one in the context of driving.

Contributions to the state of art

This problem concerns a lot of different branches of knowledge such as human emotion, ADAS, sound or Deep Learning. This specific project contributes to the state of art mainly through the followed methodology in how data is treated before the training process and the comparison between different approaches.

SYSTEM'S ARCHITECTURE

The speech recognition component is designed to be implemented inside a complete ADAS made of two other components: the microphone and the actuator system. In the first place the microphone would obtain the speech sound from the car and process it so a computer can handle it. After the audio is digitalized, the designed component processes the audio extracting its features and using the CNN model to recognize the emotion. Once the emotion is identified, it is communicated to the actuator system, which is in charge of taking action to avoid a car accident. This action could be a kind of alarm to warn the driver or the adaptation of other assistance systems for that specific emotion. All this architecture is shown in Fig. 10.3.

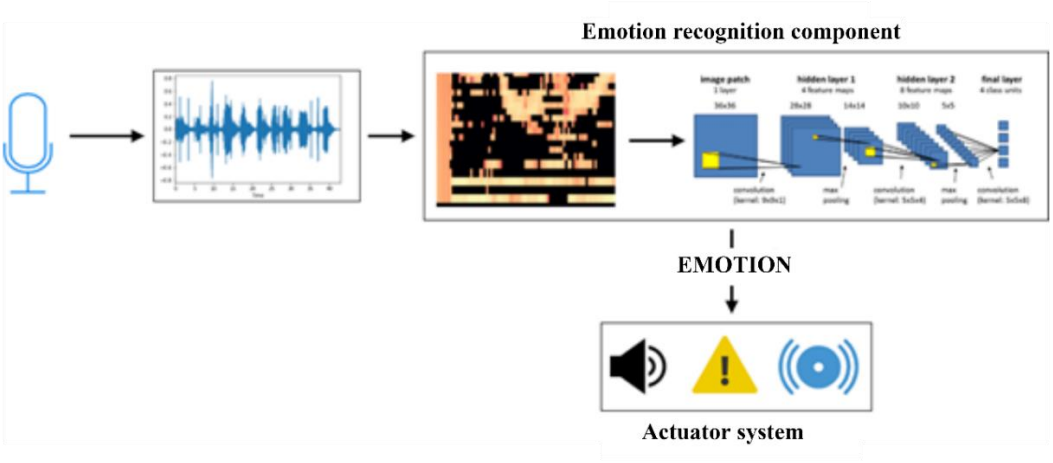


Fig. 10.3. System's architecture diagram where the designed component is integrated

DATASETS AND DATA PROCESSING

In order to create a model capable of generalizing a problem, it is necessary to obtain a dataset composed of the instances that characterize this problem. This set of instances needs to be diverse, balanced and have quality data to obtain an accurate model.

Source and data characteristics

The lack of a consensus about a set of basic emotions and the difficulty of creating a valid dataset of acted emotions makes obtaining a dataset a difficult task. In this project, there are three different data sources which contribute to the diversity of the dataset.

The Toronto Emotional Speech Set (TESS) [37] has a total of 2800 speech recordings made by two actresses of 26 and 64 years old in English. In them a total of seven emotions are performed: fear, pleasant surprise, sadness, anger, disgust, happiness and neutral. The recordings are structured and only a word is changed in them, all of them being recorded with a sampling rate of 22050 Hz. The main disadvantage of this dataset is the demographic unbalance due to the sex of the persons in the recordings.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [38] has a total of 7356 recordings of audio, video and both at the same time. In them 12 actors and 12 actresses sing and say two different predetermined phrases performing a total of eight emotions: neutral, calm, happiness, sadness, anger, fear, disgust and surprise. All the recordings have a sampling rate of 48000 Hz. Its advantages are the balance of the instances and the validation process done to verify the correctness of the emotions performed.

The Berlin Database of Emotional Speech (BERLIN) [39] has a total of 535 speech recordings where 5 actors and 5 actresses perform 10 different phrases in 6 different emotions: anger, boredom, disgust, anxiety/fear, happiness, sadness and neutral. The performers speak in German and are between 20 and 35 years old. The recordings have a sample rate of 22050 Hz, with the advantage of being balanced but the drawback of having significantly less instances than the other datasets.

Extraction and labeling

In order to be able to work with the different datasets as a whole, it is necessary to extract the data from its original format into a unified one with homogeneous labels. This process is performed by parsing each file from its directory and creating a composed database with all the instances.

The instances have different sampling rates which means that audios with the same duration have different numbers of values. This issue is solved by ignoring the duration of the audios and dividing the instances in 20.000 sample windows, which would create new instances with a duration between 0.9 and 0.4 seconds. In addition to the division of the original instances, while the datasets are extracted a function to trim the silent parts is applied so there are not irrelevant values.

While the instances are extracted from their respective directories, the corresponding label is assigned to the emotion that identifies it. This label is an integer number from 0 to 8 which corresponds to the index of the emotion in the following list of nine emotions: anger, disgust, fear/anxiety, happiness, neutral, surprise, sadness, boredom and calm.

Data processing

Once the instances are homogeneous it is necessary to process them to produce a richer representation that facilitates the learning process of the model and improves its accuracy once it is trained. This data processing is achieved by applying an audio transformation

called Mel Frequency Cepstral Coefficients (MFCC). This transformation transforms the one-dimension vector representing the amplitudes over time, into a two-dimension matrix which represents the value of the intensity depending on the frequency expressed in mels over time. The steps of the MFCC transformation are the following:

1. Segment the input into time windows.
2. Apply the Fourier Transformation to each one of the windows obtained in the previous step. This transformation gives the pure frequencies which addition produces the input's complex wave.
3. Take logarithms to each of the coefficients.
4. Apply the discrete cosine transformation to each of the results.

The mel unit in which the frequencies are expressed is obtained by applying a non-linear transformation to the values expressed in Hz. This transformation is done because the mel scale is built accordingly to the capacity of humans to differentiate different frequencies, which is more sensitive for lower values than for higher ones.

Training datasets generation

Once the data is processed, it is necessary to create the different subsets that are used in the training process. There are two big groups of datasets depending on the form of representing the data: the first one is composed of the 2D matrices that were obtained from the data processing step, and the other one is composed of the images obtained by representing those 2D matrices in the same way as a spectrogram and saving it as an image. An example of these images is shown in Fig. 10.4.

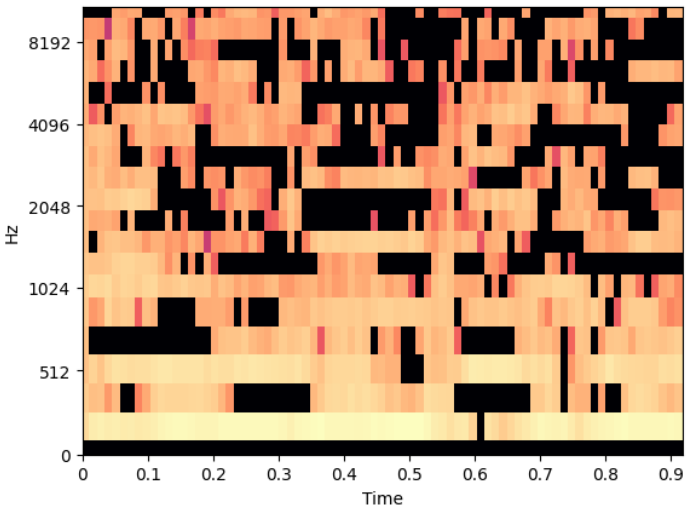


Fig. 10.4. Spectrogram representation of the MFCC transformation of an audio register

There are two main reasons behind transforming the matrices into images: the first one is the influence of window size in the representations, which is not appreciated in the

numeric representation, that makes the width of the rectangles of the image vary its size depending on the window size and its hop size. This difference is shown in Fig. 10.5.

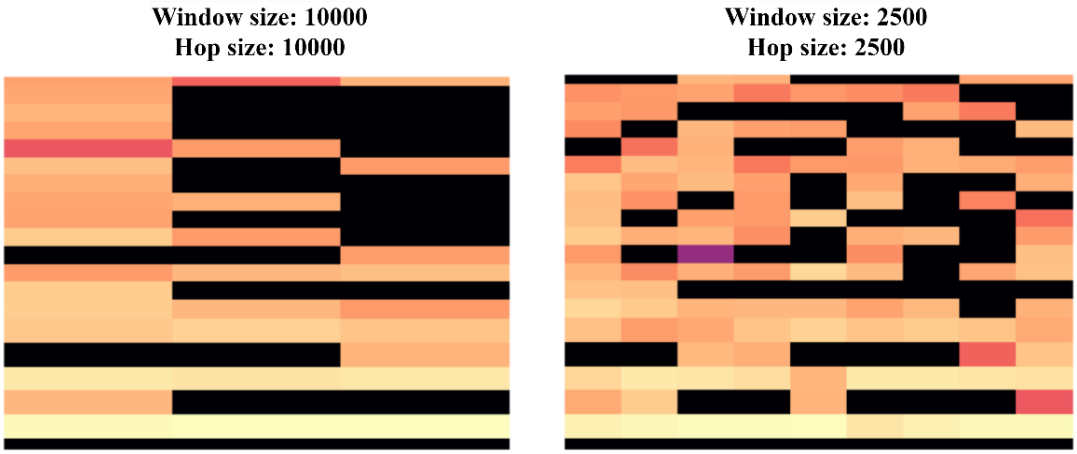


Fig. 10.5. Comparison between images generated with different window and hop size

The other reason is the influence of the mel transformation in the graphical representation that is not perceived in the numeric one. This difference is shown in Fig. 10.6, where the rectangle's height is smaller at the ones at the top than the ones at the bottom, which is due to the mel non-linear transformation.

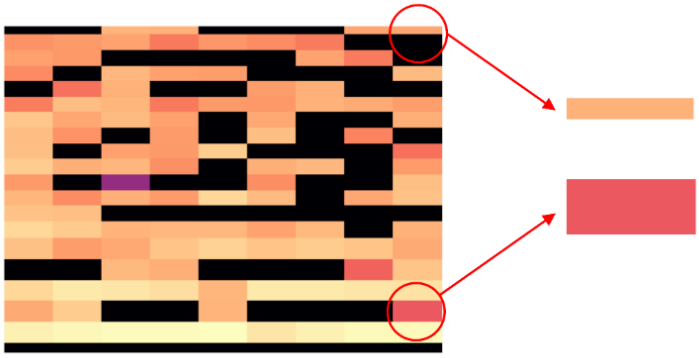


Fig. 10.6. Comparison between windows' size for high frequencies with shorter windows (orange) and low frequencies with higher windows (red)

These two groups of numeric values (M) and images (I) are each one divided in three different subsets which are created by combining the different data sources. The resulting combination of subsets are one subset containing data from the three sources combined, another with one data from RAVDESS and TESS and the last one containing data only from RAVDESS and BERLIN. These subsets were obtained through experimentation but there are reasons behind the changes done in the second and third ones. The BERLIN source was deleted in the second one in order to study if using only instances in English could improve the results and the TESS was deleted in the third subset in order to verify if the sex balancing affected the results. These data subsets are defined in TABLE 10.1.

TABLA 10.1. GENERATED DATA SUBSETS

Dataset	Number of instances	Tunnings made
RBT_I	8750	Boredom and calm classes are deleted. Subsampling that leaves 1250 instances per class.
RBT_M	8750	Boredom and calm classes are deleted. Subsampling that leaves 1250 instances per class.
RT_I	7700	Boredom class is deleted. Subsampling that leaves 1100 instances per class.
RT_M	7000	Boredom class is deleted. Subsampling that leaves 1100 instances per class.
RB_I	5040	Neutral and boredom classes are combined. Subsampling that leaves 630 instances per class.
RB_M	5040	Neutral and boredom classes are combined. Subsampling that leaves 630 instances per class.

The six different training data subsets were all balanced after creating them using subsampling because of the simplicity of applying this type of balancing technique and because the capability of doing it due the elevated number of instances.

SPEECH EMOTION RECOGNITION

Once all the training data is prepared, different CNN models are created through an experimentation process which explores different network architectures, hyperparameters and datasets to find the model that makes the best generalization of the problem.

Eight different CNN architectures are generated, four for each data approximation because such different inputs require different topologies, and each of them is trained with a different set of hyperparameters such as learning rate, number of epochs, optimizer,

batch size and loss function. The summary of this experimenting process is detailed in TABLE 10.2.

TABLA 10.2. EXPERIMENTATION PROCESS SUMMARY

Images			**Decimal Values Matrices**		
Experimenting time	**Results**		**Experimenting time**	**Results**	
	Maximum	**Media**		**Maximum**	**Mean**
15 hours y 50 minutes	93.34 %	55 %	1 hour	71.27 %	56.8 %

In order to validate some of the most relevant experiments, a technique called Cross Validation was applied, which consists in creating several random test and train partitions in order to train and test the model with them and calculate the mean of the evaluation metrics. This approach allows to verify that the results that are obtained are not biased by a specific train-test split configuration.

The final model which was selected used the image approach and the dataset without instances in German, which was negatively affecting the results of the models. This model has a **93.34%** accuracy for test instances which means that successfully generalizes the problem and its convolutional architecture is shown in Fig. 10.7.

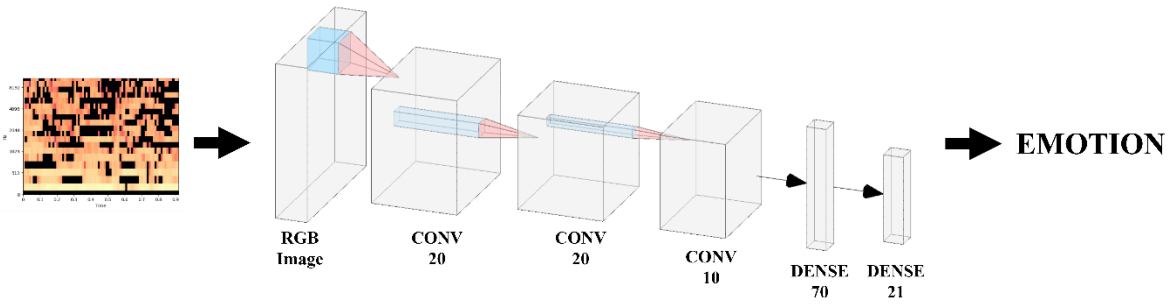


Fig. 10.7. Final model's convolutional architecture

Comparison between approaches

Although both data approximations were solved using the same technique, they have substantial differences between their results and the way to obtain them. The chosen model used images as input for training and prediction, but there are some applications where this approach could not be the best one. The image approach shows that the extra information, gained from creating the images from the matrices outputted by the MFCC transformation, allows to increase the accuracy up to 93%. However, when data in German is introduced, the matrices approach gives better results.

Another disadvantage of using images is the computational cost associated with them, whose graphical representation introduces 2 new channels for RGB coloring and multiplies the number of values representing each instance by 380. This creates a slower learning and prediction system which requires more powerful hardware and bigger

amounts of storage, which could make it not suitable for situations where hardware is limited and the required response time is minimal.

SOCIO-ECONOMIC IMPACT

New developed technologies have direct impacts in society that can be positive or negative, which can also impact the industry that developed it. This specific technology is designed to reduce negative consequences of car accidents, which is estimated to cost between 1% and 2% of a country's Gross National Product (GNP) each year [45], in addition to all the human losses.

The ADAS technology aims to reduce both losses, in addition to other issues related to car accidents like the environmental impact they produce in the ecosystem where the accident happens.

The introduction of this type of system could also benefit the industry, because having a system that could be used to adapt other ADAS depending on the emotion detected, upgrades all these types of products.

CONCLUSION

After the development of this whole project, it can be concluded that both principal and specific objectives were achieved. The State of Art has given a general overview of the situation in which the solution is designed, the experimentation phase has produced 58 valid experiments with a total of almost 16 hours of training time and finally the model has been validated to test its correctness.

The biggest difficulties found were related with human emotion, whose difficulty to make a closed definition and its variability in speech through different languages, cultures and demographic factors, make their recognition a very challenging task. This recognition is even harder when it can only be done using speech while there are many other physiological and contextual factors that define it.