UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA
División de Estudios de Posgrado

Reconocimiento híbrido de emociones a partir del análisis de voz y de expresiones faciales

TESIS
PARA OBTENER EL TÍTULO DE
MAESTRO EN TECNOLOGÍAS DE CÓMPUTO APLICADO

Presenta:
Ing. Jorge Arturo Carrasco Jiménez
Director de tesis:
Dr. Raúl Cruz Barbosa
Codirector de tesis:
Dr. Arturo Téllez Velázquez

Huajuapan de León, Oaxaca, México, mayo de 2023

Dedicado a Dios
y a las personas que me apoyaron
directa e indirectamente en este camino

Agradecimientos

A mis directores, Dr. Raúl Cruz Barbosa y Dr. Arturo Téllez Velázquez por el tiempo invertido, por la dedicación puesta en este trabajo, por sus acertados comentarios y observaciones hacia cada uno de los aspectos importantes de esta tesis.

A mis sinodales, Dr. José Anibal Arias Aguilar, Dr. Eduardo Sánchez Soto, Dr. Christian Eduardo Millán Hernández y Dr. Rosebet Miranda Luna por su generosa lectura y acertados comentarios hacia esta investigación.

A la Mtra. Yoshaira Soledad Alexandres Carrizosa, por su apoyo en la revisión de este documento.

A la Universidad Tecnológica de la Mixteca y a sus profesores, en específico a todos aquellos que contribuyeron en mi formación académica.

A la M.C. Ivette Jiménez García, por su apoyo en la validación de UTeMo.

A los grupos de teatro: Dii Ini Na, municipal de Huajuapan de León y Universidad Tecnológica de la Mixteca, por su participación en la construcción de UTeMo.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT), por el apoyo económico recibido a lo largo de este proyecto de tesis.

A mi familia, por su apoyo incondicional en toda mi vida.

A mis amigos, por todos los buenos momentos juntos.

A Mayra Agama, por creer en mí y darme ánimo todos los días.

Índice general

Índice de figuras

Resumen

Recientemente, la importancia del reconocimiento automático de emociones aumenta cada vez más en las aplicaciones con interfaces humano-computadora. Para lograr dicha interacción, las emociones juegan un papel muy importante, por lo que, saber reconocerlas agregaría mayor certeza en este proceso. Sin embargo, la tarea de reconocer emociones es desafiante, ya que el desarrollo de modelos precisos de reconocimiento involucra el estudio de otras áreas o ciencias afines, tales como las ciencias cognitivas, la psicología y las ciencias computacionales.

Debido a que la lengua influye en las expresiones emocionales a través de las emisiones de voz, en el presente trabajo de tesis se construye una base de datos audiovisual en español de México. Esta base de datos denominada UTeMo se construye con la ayuda de siete actores y siete actrices, quienes interpretan diferentes frases basadas en las siete emociones del modelo de Paul Eckman, además de un estado neutral, obteniendo un total de 1801 muestras de video. Cabe mencionar que, estas muestras son validadas de manera cualitativa y cuantitativa. La validación cualitativa se realiza mediante un análisis estadístico y una verificación por un experto que consiste en asegurar una correcta interpretación de la emoción; mientras que la validación cuantitativa se realiza utilizando algoritmos tradicionales de aprendizaje supervisado. Posterior a esta validación se trabaja con métodos de aprendizaje profundo para el análisis de voz y de expresiones faciales por separado.

Con el fin de encontrar modelos precisos de alta exactitud de clasificación para el reconocimiento de emociones mediante la voz, se utilizan diferentes técnicas con redes neuronales convolucionales 1D y 2D. En la red 2D utilizada (FSER) se propone una transformación de información de audio a imagen a través de una representación espectral. Experimentalmente se determina que, para UTeMo, se extraen las características más representativas de cada emoción a partir de los cocleogramas, ya que con esta representación se alcanza un 95.08 % de exactitud de clasificación.

En el caso del reconocimiento de emociones mediante expresiones faciales se

toma en consideración el análisis temporal entre imágenes mediante técnicas de flujo óptico. Sin embargo, experimentalmente se determina que hay un mejor desempeño realzando las características de las imágenes de los rostros y utilizando éstas como fuente de información. Utilizando las imágenes de las expresiones faciales con realce de características y la arquitectura MobileNet se logra un 95.96 % de exactitud de clasificación.

Finalmente, con el fin de mejorar la exactitud de clasificación se realiza la combinación de ambos modelos, obteniendo así un modelo híbrido que utiliza la salida de los modelos de voz y expresiones faciales para generar una nueva decisión. Al realizar esta combinación se logra un 99.26 % de exactitud de clasificación. De acuerdo a los resultados obtenidos, se verifica que el reconocimiento de emociones mejora significativamente al utilizar dos fuentes de información.

Capítulo 1

Introducción

Las emociones humanas son un fenómeno complejo y ambiguo, difícil de modelar matemáticamente, ya que dependen de diferentes factores como el género, la edad, la cultura o el lenguaje. En los últimos años, se ha tomado en consideración el desarrollo de sistemas inteligentes de reconocimiento de emociones, debido a la gran capacidad que tienen las computadoras modernas para resolver este tipo de problemas [Rahdari et al., 2019].

Es cierto que las emociones son propias de las personas, lo cual las distingue de las máquinas, es por esta razón que la investigación del reconocimiento de emociones humanas ha ganado mucho interés en el sector académico e industrial. Esto es la identificación de emociones con un alto grado de precisión puede ser muy útil para diversas aplicaciones y sistemas, tales como: servicios médicos, computación afectiva, juegos hápticos, mantenimiento seguro de ciudades inteligentes, entre otras [Kim et al., 2020].

La posibilidad de reconocer emociones humanas se ha logrado a través de diversas fuentes, ya sea mediante información de audio, información visual o señales bioeléctricas. Es así que esta tesis tiene como objetivo principal la implementación de un algoritmo híbrido bimodal que trabaje tanto con expresiones faciales como con el análisis de voz para el reconocimiento de emociones. De esta manera, mediante el uso de técnicas de aprendizaje profundo, se verifica que se mejora considerablemente la clasificación de las siete emociones básicas: ira, miedo, felicidad, tristeza, asco, sorpresa y un estado neutral.

Para lograr el propósito antes mencionado, se construye una base de datos con información audiovisual en español mexicano, con el fin de mejorar la identificación

y extracción de características en este contexto. Esta base de datos es validada tanto cualitativamente como cuantitativamente, con el fin de mostrar su fiabilidad al momento de hacer reconocimiento de emociones. También, se realiza un análisis de algoritmos de aprendizaje profundo centrado en las redes neuronales convolucionales con la finalidad de obtener una mayor discriminación entre emociones. Para el análisis de expresiones faciales se consideran un realce de características y las relaciones temporales que hay en las expresiones faciales; mientras que para el análisis de voz, se consideran diferentes métodos de extracción de características, así como la transformación de información de audio a imagen mediante representaciones espectrales.

Este documento tiene cinco capítulos: en el primer capítulo se presenta el planteamiento de la tesis. El segundo capitulo ofrece una breve descripción de los principales temas que sustentan esta tesis. El tercer capítulo presenta la metodología seguida para la construcción de la base de datos y la implementación del método para reconocer emociones. El capítulo cuarto muestra los resultados obtenidos y finalmente en el capítulo quinto se presentan las conclusiones y trabajo a futuro.

1.1. Planteamiento del problema

En los últimos años, la interacción humano computadora se ha vuelto sumamente estrecha. Sin embargo, esta relación se ve limitada porque las tecnologías modernas, llámense computadoras, dispositivos, aplicaciones, dispositivos móviles o *gadgets*, no ofrecen un reconocimiento en cuanto a expresión de emociones.

La interacción que se enfoca en las emociones del usuario mientras interactúa con la computadora y aplicaciones se le ha denominado computación afectiva [Calvo et al., 2014]. La visión de la computación afectiva es hacer que los sistemas sean capaces de reconocer las emociones humanas e influir en ellas para mejorar la productividad y la eficacia del trabajo con computadoras u otras tecnologías [Hippe et al., 2014].

Hay muchas formas diferentes de poder clasificar las emociones y son varios los investigadores que han abordado este problema. Existen enfoques que han trabajado el reconocimiento de emociones unimodalmente mediante coeficientes cepstrales de frecuencia de Mel (MFCC, por sus siglas en inglés) [Rani and Yadav, 2018], mientras que en otras investigaciones han trabajado con señales electroencefalográficas [Liu et al., 2011].

El reconocimiento de emociones desde un enfoque multimodal parte de las ex-

presiones faciales y el ritmo cardiaco (Du et al., 2020), así como de la extracción de características de las expresiones faciales y del audio [López Gil and Garay Vitoria, 2019]. Otra clasificación de las emociones involucra las expresiones faciales usando información de imagen, se ha propuesto el uso de Redes Neuronales Recurrentes y de Redes Neuronales Convolucionales (RNN y CNN respectivamente, por sus siglas en inglés) [Zhang et al., 2018]. Además, para este mismo propósito, pero usando información de audio, se han utilizado CNNs y Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) [Fan et al., 2016]. Otra idea interesante planteada en la literatura es que el audio se trabaje como imagen, utilizando una representación bidimensional del espectro en frecuencia de una señal de audio [Ramírez Cornejo and Pedrini, 2019].

En esta tesis se plantea realizar un reconocimiento de emociones a partir del análisis de voz en conjunción con las expresiones faciales, con el fin de mejorar la discriminación en la clasificación de las emociones básicas: tristeza, felicidad, ira, asco, sorpresa, miedo y un estado neutral. Para poder lograr este objetivo, se realiza un preprocesamiento tanto en audio como de imagen que permite una mejor extracción de características, de tal manera que se pueda realizar un análisis de rendimiento eficaz entre la presente propuesta y algunos otros enfoques multimodales. Es así que se realizará una selección de características apropiada y un análisis cuantitativo de algoritmos de aprendizaje supervisado empleados para realizar esta difícil tarea.

1.2. Justificación

La emoción es un término general empleado para una serie de experiencias cognitivas subjetivas. Las emociones consisten en un conjunto de estados psicológicos generados por diversos sentimientos, pensamientos y comportamientos. Generalmente, la gente transmite información emocional constantemente durante el proceso de comunicación [Davidson et al., 2018].

Por lo anterior, el reconocimiento de emociones juega un papel importante en la comunicación interpersonal y en muchos aspectos de la vida diaria. Por ejemplo, reconocer los estados emocionales de los pacientes con trastornos de la expresión emocional sería útil para proporcionar mejores tratamientos y cuidados. Asimismo, también es un aspecto indispensable de la humanización en la interacción humano-computadora [He et al., 2020].

Dado que las emociones juegan un papel importante en la vida diaria de los seres humanos, la necesidad y la importancia del reconocimiento automático de emociones

aumenta cada vez más en las aplicaciones con interfaces humano-computadora [Liu et al., 2011]. Por ejemplo, [Liu et al., 2011] desarrollaron prototipos basados en señales electroencefalográficas para musicoterapia y una sugerencia de canciones en un sitio web, de acuerdo con el estado de ánimo detectado. Además, afirman que hoy en día con los medios digitales, las nuevas formas de interacción entre el humano y la computadora tienen el potencial de revolucionar el entretenimiento, el aprendizaje y muchos otros aspectos de la vida humana.

Otra aplicación del reconocimiento de emociones se encuentra en la tecnología móvil. Debido al crecimiento continuo del uso extensivo de teléfonos inteligentes, servicios y aplicaciones, el reconocimiento de emociones se está convirtiendo en una parte esencial para brindar atención emocional a las personas. El brindar atención emocional puede mejorar enormemente la salud y experiencia de los usuarios. Además, debido a la dinámica y heterogeneidad de las aplicaciones y servicios móviles, es un desafío proporcionar un sistema de reconocimiento de emociones que pueda recopilar, analizar y procesar comunicaciones emocionales en tiempo real y de manera altamente precisa, con un tiempo de procesamiento mínimo [Hossain and Muhammad, 2017].

Debido a la naturaleza compleja de las emociones humanas, el reconocimiento automático de dichas emociones sigue siendo una tarea desafiante. Una dificultad que tiene que afrontar una máquina es el hecho de que los seres humanos rara vez expresan sus emociones de forma exclusiva, utilizan varios canales, como el habla y la mímica [Wagner et al., 2011]. Debido a esto, diversos autores [Du et al., 2020, López Gil and Garay Vitoria, 2019] se han enfocado en el reconocimiento multimodal de emociones, es decir, reconocer emociones a partir de diferentes fuentes que se complementan, por ejemplo, de la voz y expresiones faciales. Para el caso de la voz, se han utilizado diferentes idiomas como el inglés o alemán, pero hasta la fecha no se ha encontrado ninguna publicación asociada con el español mexicano. Por lo anterior, uno de los objetivos del presente trabajo es recolectar un conjunto de datos con audio en español mexicano.

La tarea de reconocimiento de emociones de este trabajo de tesis se realiza de manera híbrida, haciendo un análisis de voz y de expresiones faciales. Para esto, el análisis de audio se procesa a partir de sus características unidimensionales y de su representación bidimensional espectral, para extraer de éstas las características con una CNN, tal como se realiza convencionalmente con las imágenes. Por otro lado, el análisis de las expresiones faciales se hará con un proceso similar, aplicando un preprocesado para resaltar determinadas características del rostro. Finalmente las decisiones obtenidas de cada clasificador se combinan mediante una regla de decisión. Debido a que se procesa un gran número de imágenes, se justifica el uso de las herramientas de cómputo paralelo, tal como lo es la Unidad de Procesamiento

Gráfico (GPU, por sus siglás en inglés),

1.3. Hipótesis

Por medio del reconocimiento de emociones multimodal, a partir del análisis de la voz y de las expresiones faciales, es posible obtener una mejora en el desempeño respecto a exactitud de clasificación de las emociones básicas en el español mexicano, a través del aprendizaje profundo.

1.4. Objetivos

1.4.1. Objetivo general

Analizar e implementar un algoritmo de reconocimiento híbrido de emociones a partir del análisis de la voz y de las expresiones faciales, con la finalidad de mejorar el desempeño de clasificación de las emociones básicas mediante técnicas de aprendizaje profundo.

1.4.2. Objetivos específicos

1. Revisar el estado del arte sobre los métodos utilizados en el reconocimiento de emociones a partir de la voz y de las expresiones faciales, así como las técnicas utilizadas para el preprocesamiento de audio e imagen.

2. Construir un conjunto de datos de audio en español mexicano y expresiones faciales que contenga las seis emociones básicas: ira, tristeza, felicidad, miedo, asco, sorpresa y un estado neutral.

3. Seleccionar los métodos de preprocesamiento de audio e imagen, que ayuden a realzar las características principales presentes en la voz y las imágenes.

4. Implementar un algoritmo de reconocimiento de emociones bimodal, vía voz y expresiones faciales que utilice técnicas de aprendizaje profundo.

5. Validar el rendimiento del algoritmo de reconocimiento de emociones, usando medidas de rendimiento para aprendizaje supervisado.

6. Realizar un análisis comparativo de los resultados de la presente propuesta con al menos una del estado del arte.

1.5. Metas

1. Construcción de una base de datos audiovisual en español mexicano.

2. Elaboración de un reporte de algoritmos de preprocesamiento de audio e imagen para el reconocimiento de emociones.

3. Implementación de un módulo de reconocimiento de emociones multimodal para mejorar la discriminación de emociones.

4. Creación de una biblioteca en lenguaje `Python` para ejecución de los algoritmos seleccionados.

5. Elaboración de un cuadro comparativo de rendimiento del algoritmo implementado y otras soluciones encontradas en el estado del arte.

6. Publicación de un artículo.

1.6. Limitaciones

El presente trabajo de tesis se limita a realizar la identificación únicamente de las seis emociones básicas: ira, tristeza, sorpresa, felicidad, miedo y asco más un estado neutral, mediante un algoritmo de aprendizaje supervisado como pueden ser las CNN's. La base de datos se limita a contener información audiovisual, ya que este trabajo hará uso de la información en formato de imagen y audio, a su vez se consideran únicamente las emociones de: ira, asco, felicidad, tristeza, sorpresa, miedo y un estado neutral.

1.7. Trabajo relacionado

El estudio de las emociones ha tenido diferentes enfoques a lo largo de la historia por lo que han surgido diferentes modelos con el fin de encontrar una clasificación e identificación de estas, siendo las principales corrientes: biológica, conductual, cognitiva y social [Bisquerra Alzina, 2009]. Debido a que la corriente biológica se centra en las expresiones faciales y la universalidad de las emociones, se le prestará mayor interés para fines de este escrito.

La tradición biológica se inicia con Darwin, quien pensaba que las emociones comunican intenciones que tienden a ser reacciones apropiadas a ciertos acontecimientos del entorno [Darwin, 1872]. A partir de esta corriente, diversos autores fueron construyendo sus teorías (por ejemplo: Omkins, Ekman, Izard, Plutchik y Zajonc), siendo resaltable la clasificación emocional de Paul Ekman, cuyo modelo emocional basado en expresiones faciales [Ekman, 1992] ha sido utilizado para construir diversas bases de datos.

De manera general, los postulados neodarwinistas acerca de las emociones pueden resumirse en que son reacciones universales, adaptivas, heredadas y desarrolladas conforme hay maduración biológica [Palmero et al., 2002]. Además, tienen formas de expresión facial y corporales determinadas, por lo cual, existen emociones básicas.

De manera más específica, tal y como Ekman lo menciona en [Ekman et al., 1983] desarrollaron un procedimiento de análisis de los movimientos de los músculos faciales, denominados FACTS (por sus siglas en inglés, *Facial Action Coding System*) para determinar una clasificación de emociones. Cada emoción tiene un patrón transcultural de expresión facial; existen patrones neurofisiológicos universales y específicos asociados a cada emoción básica; implica múltiples señales, tanto faciales como vocales; ciertas asociaciones entre estímulo y emoción pueden estar medidas por el aprendizaje y cada expresión es breve (dura de 0.5 a 4 segundos).

Así como desde el punto de vista humano se habla de la empatía, que es la capacidad mental y afectiva de identificar el estado emocional de otra persona [Moya Albiol, 2014], también se han buscado formas computacionales para reproducir esta habilidad, que normalmente se le denomina reconocimiento de emociones. Para el reconocimiento de emociones se ha trabajado desde diferentes enfoques, tanto unimodales como multimodales; estas palabras hacen referencia a la cantidad de fuentes de información que se toman como puntos de partida para el entrenamiento de los diferentes algoritmos y sistemas de clasificación.

La información con la cual se hace el reconocimiento de emociones puede ser

variada, por ejemplo: se puede utilizar el ritmo cardiaco [Du et al., 2020], expresiones faciales [Ozdemir et al., 2019, Liu et al., 2017], voz [Rani and Yadav, 2018], señales biológicas [Liu et al., 2011, Shin et al., 2018, Kołakowska et al., 2014], entre otras fuentes. Un ejemplo unimodal es la arquitectura que fue propuesta por [Wootaek et al., 2016] para abordar el problema del reconocimiento de emociones en el habla (speech). En ésta se propone una unión entre una CNN (Red neuronal convolucional) y una arquitectura LSTM; mientras que un caso multimodal puede ser la información en conjunto entre el ritmo cardíaco y la información de las expresiones faciales para obtener una mejora en el desarrollo de los juegos de video. Esta mejora consiste en maximizar la experiencia de un jugador trabajando con el reconocimiento de emociones a partir de ambas fuentes de información, capturando el ritmo cardiaco con un equipo especializado, por un lado y las expresiones faciales con un Kinect, por otro; esta clasificación fue hecha de manera independiente utilizando inteligencia artificial, redes neuronales convolucionales para las expresiones faciales y bloques LSTM (por sus siglas en inglés *Long Short Term Memory*) para el ritmo cardiaco. A su vez, estas fueron fusionadas con redes de mapas auto organizados lo cual logra el reconocimiento de la emoción total [Du et al., 2020].

Dentro de las señales biológicas, las más populares son las señales electromiográficas (provenientes del sistema nervioso-muscular) y electroencefalográficas (que provienen de la corteza cerebral) [Liu et al., 2011, Kołakowska et al., 2014]. Existe un caso en particular que incluye la unión de estas dos fuentes de información y adicionalmente las señales de electro-oculograma que son señales provenientes de los músculos de los ojos. La arquitectura de clasificación utilizada se enfoca en reconocer 5 estados emocionales (basados en la teoría de Eckman) y el algoritmo que se utiliza son las máquinas de soporte vectorial [Shin et al., 2018].

Para el caso particular de la voz, existen diversas características a extraer. Hay trabajos que se centran en clasificar y ver qué características de audio pueden ser efectivas para hacer clasificación de emociones [Pérez Espinosa and Reyes García, 2010, Xu et al., 2014]; mientras que otros se centran en la tarea de clasificación, cuyo recurso usualmente utilizado es el habla [Reddy and Kuchibhotla, 2019] y tono [Nerio et al., 2018, Chandrasekar et al., 2014]. A su vez, se utilizan las máquinas de soporte vectorial. De igual forma, hay propuestas que trabajan con información de audio haciendo una transformación al espectro en frecuencia de la señal de voz [Rani and Yadav, 2018], un ejemplo de este tipo es el trabajo realizado por [Dossou and Gbenou, 2021], en el que se obtiene un resultado de exactitud de clasificación del 95 %. Este autor utiliza 4 bases de datos distintas: Emovo, Emodb, Ravdess y Savee que se combinan en una sola para poder clasificar las emociones.

En el caso del uso de imágenes, los algoritmos más utilizados de clasificación son las redes neuronales convolucionales, las cuales ubican las regiones de interés

principales en los ojos, nariz y boca. Las emociones que las bases de datos incluyen normalmente son las del modelo emocional de Paul Eckman. Cabe resaltar que existen trabajos que mencionan el uso de técnicas de aprendizaje computacional extremo (por sus siglas en inglés extreme learning Machine) [Ozdemir et al., 2019, Liu et al., 2017].

Cuando se trabaja con información audiovisual, en la mayoría de casos, se aborda desde dos enfoques. El primer enfoque combina las características de audio y de imagen (visuales) en un solo arreglo, el cual será procesado por el algoritmo de clasificación. Ejemplo de esto es el *framework ISLA* (Enfoque de Segmentación y Etiquetado Informado, por sus siglas en inglés) que consiste en procesar la información audiovisual como si fuera un solo vector [Kim and Provost, 2019].

El segundo enfoque, procesa por separado las características visuales y auditivas, para posteriormente hacer la integración de la información y llegar a una conclusión [Avots et al., 2019]. Ejemplo del segundo enfoque es [Hossain and Muhammad, 2019] un sistema desarrollado que propone una CNN en 2D para analizar el *Speech* del audio y una CNN 3D para las imágenes provenientes de un clip de video. Posterior al proceso individual de la información, las salidas de cada red se conectan a una red tipo *feed fordward* y las salidas probabilísticas que ésta produce sirven como entrada para un clasificador basado en máquinas de soporte vectorial.

Una forma con la cual se ha trabajado el audio en información audiovisual, es mediante el cambio de la información, es decir, se pasa el audio a una imagen. La transformación de la imagen se obtiene a partir de las diferentes representaciones que se extraen del audio, por ejemplo a través de su representación espectral. En un caso en particular, se han utilizado los espectrogramas logarítmicos de Mel en conjunto con imágenes seccionadas del rostro que previamente fueron trabajadas a través de un histograma adaptativo de contraste limitado y su transformada Census. Todas estas fuentes de información se procesaron con una CNN con el fin de obtener un clasificador de emociones [Ramírez Cornejo and Pedrini, 2019].

Todos estas investigaciones han servido como guía para la realización de este trabajo, en específico aquellas que usan una base de datos de audio transformado, las que usan una clasificación por medio de redes neuronales convolucionales y los espectrogramas logarítmicos.

De los sistemas propuestos para clasificación de emociones a partir de información audiovisual, se destaca el uso de la inteligencia artificial y las diferentes formas con las que se ha abordado el problema. Una de las maneras recientes que han mostrado buenos resultados en la literatura es hacer una transformación de información para que el audio pueda representarse como una imagen y así utilizar únicamente

clasificadores diseñados para imágenes. La transformación de información más utilizada en el reconocimiento de emociones a partir de la voz es el espectrograma de Mel y de Fourier [Rani and Yadav, 2018, Hossain and Muhammad, 2019]; sin embargo hay otras representaciones espectrales utilizadas en análisis de audio (de manera general) como los cocleogramas [Cicres, 2009, Pérez Espinosa and Reyes García, 2010], con los cuales es posible experimentar y llegar a resultados similares o mejores de los ya registrados en la literatura. Para la parte visual, hay trabajos basados en una imagen de la expresión facial de la emoción; mientras que otros tratan de relacionar varios cuadros de un video [Moolchandani et al., 2021, Dino and Abdulrazzaq, 2019]. Una buena opción que se ha utilizado es la implementación de algoritmo de detección de movimiento como flujo óptico para destacar el movimiento que tiene la cara en las distintas emociones a clasificar [Anvita et al., 2020, Dino et al., 2020]. Cabe mencionar que, las regiones de interés principales son ojos, nariz y boca.

En este trabajo de tesis se busca combinar uno de estos algoritmos de detección de movimiento con una representación espectral para obtener una mejora en exactitud de clasificación respecto a otras propuestas de la literatura, así como la implementación de este método para una base de datos en español mexicano, con el fin de favorecer a la investigación local.

1.8. Metodología

La metodología que se seguirá en esta tesis consta de tres etapas principales, con el objetivo de identificar automáticamente las seis emociones básicas planteadas desde el inicio: miedo, felicidad, tristeza, ira, asco y sorpresa, además de un estado neutral [Ekman, 1992], aunado al uso de un algoritmo híbrido que funciona a partir del análisis de expresiones faciales y voz.

La primera etapa de esta investigación está dividida en dos subetapas: la adquisición de los datos y el preprocesamiento de los mismos; la segunda etapa consta de tres subfases que corresponden a la generación de características, la clasificación de las mismas y el aceleramiento; la última fase consta únicamente de la comparación de los resultados. Estas etapas de la metodología se muestran con más detalle en la Fig. 1.1.

Las especificaciones de cada etapa se mencionan a continuación:

- *Etapa 1.* En esta etapa se prepara la información para poder realizar el reconocimiento de emociones.

1. *Adquisición de datos.* Se recolectará una base de datos en español mexicano que contenga las emociones básicas y el mayor número de muestras. Asimismo, se integrará una base de datos audiovisual, utilizando imágenes y pistas de audio. Las expresiones faciales en conjunción con la voz servirán de punto de referencia para lograr el reconocimiento de emociones.

2. *Preprocesamiento.* Es esta subfase se usarán diferentes algoritmos previamente empleados y señalados en la literatura. Para el audio, se usará eliminación de ruidos y silencios, además de una amplificación de sonidos; para la imagen se utilizará ecualización CLAHE y el algoritmo Viola-Jones. A partir de los resultados, se realizará un preprocesamiento de la información, con el fin de resaltar las características principales presentes en la voz y las imágenes.

- *Etapa 2.* En esta etapa se realizará el desarrollo del algoritmo para reconocer emociones.

    1. *Extracción y selección de características.* En esta etapa se busca identificar todas las características presentes, tanto en el audio como en la imagen, que describan, especifiquen, precisen o hagan diferencia entre las emociones seleccionadas. Posteriormente, se elegirán las características más representativas de cada una de las emociones que serán usadas para esta aplicación, con el fin de resaltarlas con la finalidad de hacer más eficiente el proceso de reconocimiento de las emociones.

    2. *Clasificación.* En esta etapa se entrenará una red neuronal con la información filtrada de los pasos anteriores; se buscará entonces una configuración adecuada para la aplicación. De igual forma, se harán pruebas con otros métodos de clasificación, con el fin de encontrar el método que muestre el mejor rendimiento.

    3. *Aceleramiento.* La red neuronal se entrenará e implementará de dos formas; primera, de manera secuencial y segundo, con cómputo paralelo, utilizando GPU's.

- *Etapa 3.* Esta etapa constituye la evaluación y validación de los resultados.

    1. *Comparación 1.* Una vez implementada la propuesta, se compararán los resultados de esta investigación con otras propuestas de la literatura que fueron previamente señaladas.

    2. *Comparación 2.* Se realizará una comparación de rendimiento entre la

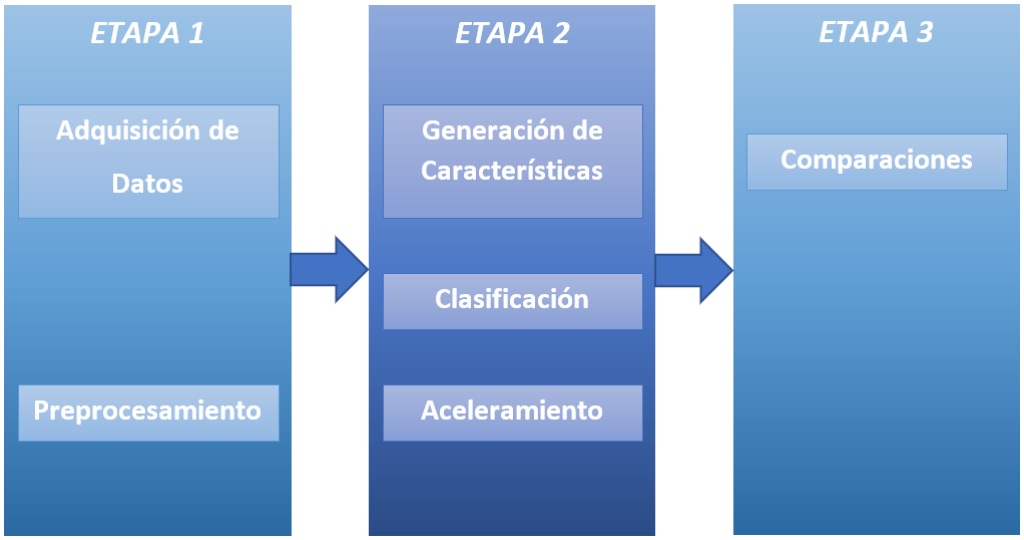


Figura 1.1: Metodología para realizar el reconocimiento de emociones

implementación paralela de la propuesta de clasificador y su contraparte secuencial.

Capítulo 2

Marco Teórico

Este capítulo ofrece una breve descripción de los principales temas que sustentan esta tesis. En la primera sección, se abordan temas relacionados con las emociones, partiendo del punto de vista psicológico. En la segunda sección, se presenta una pequeña introducción al procesamiento digital de señales para extraer características de los rostros. En la tercera sección, el enfoque principal es el audio y sus características. En la cuarta sección, se habla del reconocimiento de emociones utilizando información audiovisual. Finalmente, en la quinta sección, se habla del aprendizaje profundo y se describen las arquitecturas de redes comunes así como las utilizadas en este trabajo.

2.1. Reconocimiento de emociones básicas

La principal palabra que compete a esta tesis es *emoción*, sin embargo, no se ha definido todavía. Si se limita a la estricta etimología de la palabra, emoción quiere decir en esencia, movimiento. Es decir, expresión motora hecha a través de la conducta, ya sea por lenguaje verbal o simplemente corporal [Mora Teruel, 2013]. William James, ya en 1884, al preguntarle qué era una emoción contestó que era "una respuesta del organismo ante determinados estímulos del medio ambiente" [Mora Teruel, 2013].

Definiciones de emoción

En el Diccionario de Neurociencia de [Mora Teruel and Sanguinetti, 2004], se

entiende por emoción: "toda reacción conductual y subjetiva producida por una información proveniente del mundo externo o interno (memoria) del individuo que se acompaña de fenómenos neurovegetativos. El sistema límbico es parte importante del cerebro relacionado con la elaboración de las conductas emocionales".

[Delgado and Mora, 1998] ha definido la emoción de un modo complementario al señalar que el concepto de emoción tiene dos acepciones. En primer lugar, se puede considerar como un fenómeno interno, personalizado y difícil de comunicar a otros miembros de la misma especie. En segundo lugar, la emoción se expresa como un fenómeno externo, conductual, que sirve de clave o señal a miembros de la misma especie o de aquellos con los que mantiene una relación".

Para [Rolls, 1999] las emociones son parte de un sistema (cerebral) que ayuda a distinguir cierta clase de estímulos, muy ampliamente identificados como estímulos recompensantes o de castigo y que sirven para actuar en el mundo. Este sistema proporciona o sirve de interfaz entre tales estímulos y las conductas correspondientes.

Una definición más la proporciona [Damasio, 1999] quien menciona que las emociones son un conjunto de respuestas conducentes a mantener la vida de un organismo, las cuales implican una colección complicada de respuestas químicas y nerviosas, procesos biológicos determinados que dependen de mecanismos cerebrales innatos a pesar de que el aprendizaje y la cultura cambien.

Los mecanismos que producen las emociones ocupan un conjunto de regiones subcorticales cerebrales que engloban desde el tronco del encéfalo hasta las partes más altas del cerebro como la propia corteza cerebral. Todos los mecanismos de la emoción pueden funcionar sin deliberación consciente y todas las emociones afectan la forma de funcionar de numerosos circuitos cerebrales y el cuerpo (medio interno, visceral y sistema músculo-esquelético principalmente).

Emociones Básicas

Según [Izard, 1991], los requisitos que debe cumplir cualquier emoción para ser considerada como básica son los siguientes:

- Tener un sustrato neural específico y distintivo.
- Tener una expresión o configuración facial específica y distintiva.
- Poseer sentimientos específicos y distintivos.

- Derivar de procesos biológicos evolutivos.

- manifestar propiedades motivacionales y organizativas de funciones adaptativas.

De acuerdo con Paul Eckam existen seis emociones básicas más un estado (emoción) neutral [Ekman, 1992]:

- **Miedo:** Se identifica con el incremento de frecuencia cardíaca y la liberación de adrenalina. En cuanto a la expresión facial, el miedo se asocia con la elevación de las cejas y de los párpados superiores, la retracción de los labios y la tensión de los párpados inferiores.

- **Ira:** Como sucede con el miedo, la ira se relaciona con la activación del sistema nervioso y la liberación de adrenalina y noradrenalina; por tanto, también se identifica con el incremento de las frecuencias cardíaca y respiratoria. Según los estudios de Eckman, las cejas se acercan y descienden, mientras que los labios se aprietan. Además, aparece un "brillo" en la mirada.

- **Tristeza:** Esta emoción frecuentemente se acompaña de la disminución de la actividad motora y la aparición del llanto. La expresión facial, característica de la tristeza, se describe con el descenso de los párpados superiores y de los extremos de los labios. También se observa una menor focalización de la mirada en el punto de atención.

- **Felicidad:** La felicidad es un estado emocional agradable. Los movimientos faciales fundamentales de la felicidad son la elevación de las mejillas y la aparición de arrugas en la comisura de los ojos.

- **Sorpresa:** A diferencia del resto de emociones básicas, la sorpresa no es considerada positiva ni negativa (es decir, agradable o desagradable) sino que puede incluir componentes fisiológicos propios tanto de la felicidad como del miedo. La expresión facial de la sorpresa consiste en la apertura de la boca y de los ojos junto con la elevación de la musculatura asociada a las cejas.

- **Asco:** El asco es la emoción que expresa el rechazo a estímulos determinados que resultan desagradables para alguno de los sentidos. Podría distinguirse por arrugar la nariz y levantar el labio superior [Sasaki, 1993].

En la Figura 2.1, se muestra el resultado de un estudio realizado por Matsumoto en el cual trataron de identificar las expresiones faciales típicas de las emociones básicas. Cabe resaltar que en este estudio se toma en cuenta el desprecio como una

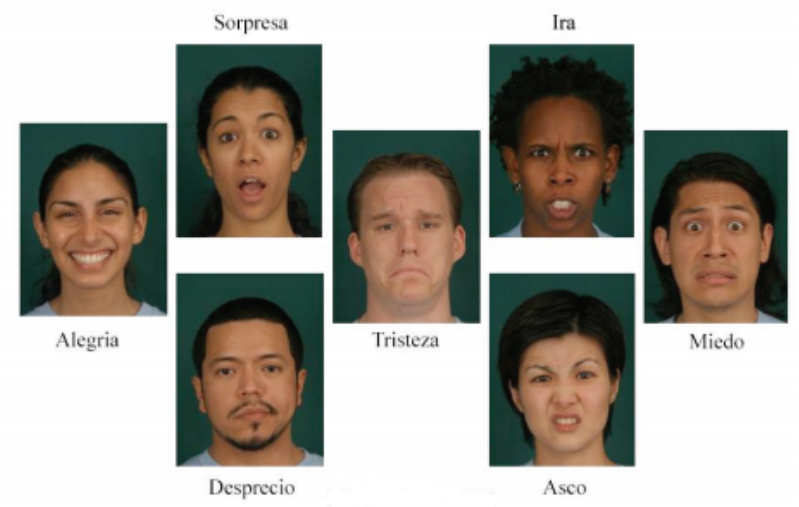


Figura 2.1: Expresiones faciales de las emociones básica según Matsumoto [Matsumoto et al., 2013], el autor incluye el desprecio como emoción básica en su investigación

emoción básica.

Las emociones desde el punto de vista psico-fisiológico

Ningún modelo de clasificación de emociones es absoluto, por lo cual se han propuesto modelos diferentes. Dichos modelos coinciden en que determinadas emociones son básicas, por ejemplo: el miedo, la ira y la tristeza que son conocidas como "las tres grandes" [Bisquerra Alzina, 2009]. El modelo de Paul Eckman está centrado en las expresiones faciales [Ekman, 1992]; sin embargo, existen cambios y reacciones neuro-fisiológicos al momento de hablar de emociones.

Uno de los primeros modelos que aparecieron fue el de [Plutchik, 1962]. Este modelo, enfocado en la adaptación biológica,indica que las emociones son reacciones de un organismo a los problemas de la vida, como mejora de la adaptación y que se presentan en estructuras de pares opuestos [Plutchik, 1962]. De igual forma, este autor introduce el concepto de intensidad emocional, ya que define los grados en las emociones (ver Figura 2.2).

En la Figura 2.2 se puede observar que entre más céntrica sea la emoción, hay una mayor intensidad. El nivel medio es el estándar promedio y el nivel exterior son emociones con bajo grado de intensidad. A su vez, hay combinaciones entre emociones, las cuales son representadas en los puntos medios.

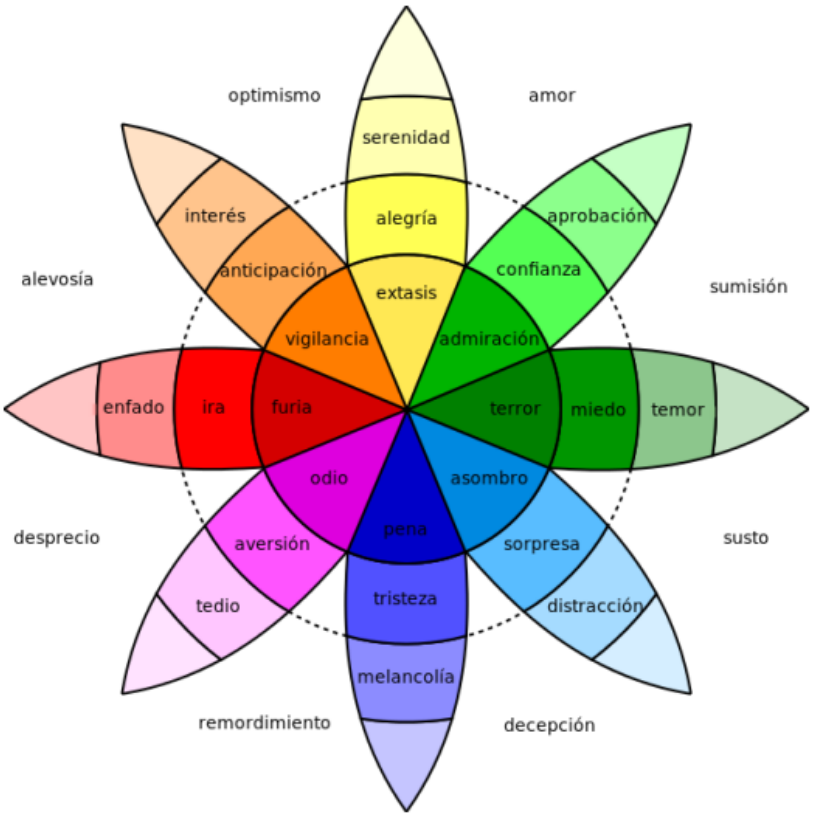


Figura 2.2: Rueda de Plutchik [Bisquerra Alzina, 2009]

Así como las emociones son reacciones corporales a estímulos externos, también estas reacciones difieren mucho entre personas, porque al estar inmersas en grupos sociales estas tienden a normalizar sus formas de respuesta, por lo cual estas reacciones se van adquiriendo de acuerdo con el condicionamiento que las personas experimentan de la cultura y la época, es decir, van cambiando. También, las emociones incluyen reacciones estructuradas heredadas, en otras palabras, las emociones incluyen conductas aprendidas [Tortosa Gil and Mayor Martínez, 1992], ejemplo de esto es que los hombres no pueden llorar o el tomar todo con humor.

Generalmente, dentro de las teorías cognitivas que existen, se piensa que para que surja una emoción debe haber una actividad cognitiva. Esta actividad va relacionada con la cualidad emocional que define el tipo de reacción o emoción que se está experimentando, así como la intensidad emocional que hace referencia al grado de activación fisiológica o grado de expresión que acompaña la reacción emocional. Por lo general, las teorías o modelos cognitivos de la emoción consideran los siguientes elementos [Fernández-Abascal and Cano-Vindel, 1995]:

- **Situación o estímulo:** Se refiere a los sucesos externos al individuo y a como los interpreta él mismo.

- **Sistema cognitivo-subjetivo:** El nivel cognitivo procesa la información, situación o estímulo, mientras que el subjetivo experimenta sensaciones, sentimientos y afectos.

- **Sistema fisiológico:** En este sistema, se desarrolla toda una serie de reacciones bioquímicas y neurofisiológicas que pueden o no ser perceptibles por el individuo.

- **Sistema expresivo-motor:** Constituye el conjunto de respuestas, principalmente de carácter motor, que emite el individuo durante este proceso. Por ejemplo, expresión facial, postura, gesto o voz.

Identificación artificial de emociones

La capacidad de hacer reconocimiento de emociones entre personas se relaciona con el concepto de empatía. Dicho término proviene del griego, sus raíces significan "dentro de él" o "lo que se siente" [Elliott et al., 2011]. Cabe mencionar que la empatía es una habilidad de la inteligencia emocional, la cual se busca integrar en las computadoras. Por lo anterior, diversos investigadores han abordado este problema desde diferentes disciplinas del conocimiento, cada uno pretendiendo encontrar un resultado óptimo [Liu et al., 2011].

Desde la disciplina del cómputo afectivo para reconocer emociones, hay investigaciones que se han enfocado en imitar el funcionamiento del cuerpo humano. Uno de los avances es el imitar la función de la amígdala [Gong et al., 2019], cuya función se relaciona directamente con las reacciones internas del cuerpo en estado de alerta. La amígdala reacciona antes que el cerebro en un determinado estado de emergencia. Para imitar este órgano, el autor presenta dos módulos que emulan la forma de procesar una emoción a través del cerebro y de la amígdala, los cuales aportan información en diferentes instantes de tiempo y con distinta magnitud.

A pesar de los avances en la imitación de partes del cuerpo, la mayoría de autores se centra en el uso de algoritmos de aprendizaje supervisado, principalmente redes neuronales [Liu et al., 2011, Du et al., 2020]. Estos algoritmos se entrenan a partir de la extracción de características de diversas fuentes de información, teniendo como principales las siguientes:

- Expresiones Faciales: Estas son extraídas a partir de imágenes o video para

analizar los gestos que muestra la persona, el cual es un recurso bastante utilizado, ya que la obtención de esta información es relativamente sencilla. En este tipo de información se hace la distinción entre si fue obtenida a partir de reacciones naturales, actuadas o avocadas [Nerio et al., 2018].

- Análisis de Voz: El atributo principalmente utilizado en este tipo de información es el *pitch* o tono; sin embargo, se pueden extraer características prosódicas, espectrales y de calidad de la voz [Perez Rosas et al., 2013].

- Señales Biológicas: Este tipo de señales se adquieren del cuerpo humano. Dentro de las más utilizadas se encuentran el ritmo cardíaco, las señales electroencefalográficas y señales electromiográficas. Las señales electroencefalográficas se obtienen de la actividad del cerebro, mientras que las electromiográficas de los músculos del cuerpo [Kołakowska et al., 2014].

Cabe resaltar que estas fuentes de información pueden utilizarse de manera independiente o combinadas.

2.2. Reconocimiento de emociones utilizando expresiones faciales

Una de las modalidades más comunes para el reconocimiento de emociones es a través de las expresiones faciales, debido a que el rostro humano es considerado el principal sistema de señales para mostrar las emociones [Alvarez and Guevara, 2009]. De manera general, el reconocimiento de emociones a partir del análisis de expresiones faciales, sigue la metodología de la Fig. 2.3 [Moolchandani et al., 2021, Revina and Emmanuel, 2021, Dino and Abdulrazzaq, 2019], la cual consiste en una entrada que consta de imágenes del rostro o secuencias de imágenes, las cuales son procesadas por algún método que ayuda a mejorar la extracción de características de las imágenes. Dichas características se ingresan al clasificador para obtener una decisión de la emoción correspondiente a la imagen de entrada. En el caso de las CNNs, el proceso de extracción de características y clasificación se puede realizar simultáneamente.

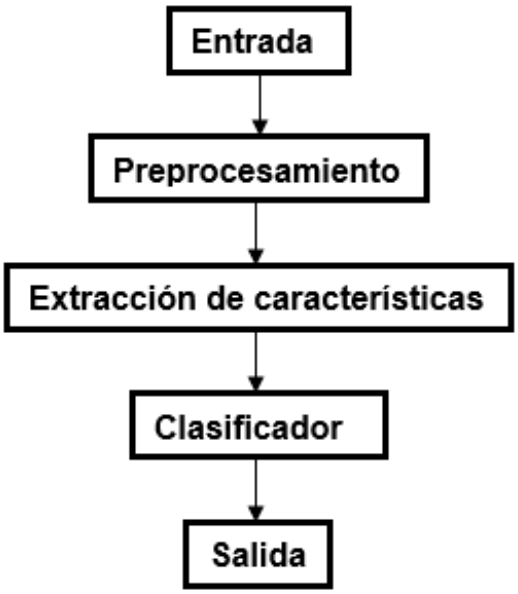


Figura 2.3: Metodología para el reconocimiento de emociones en expresiones faciales

2.2.1. Preprocesamiento digital de imágenes de expresiones faciales

El preprocesamiento se realiza con el fin de resaltar cierta información de interés y atenuar información irrelevante para la aplicación mediante la supresión de ruido, el mejoramiento de contraste, la eliminación de efectos no deseados, las transformaciones de color o algún otro método [Nixon and Aguado, 2012]. Los métodos de preprocesamiento se clasifican en tres categorías: primera, las operaciones puntuales que hacen una transformación de intensidad manipulando la intensidad de cada píxel de la imagen; segunda, las operaciones locales que operan sobre las vecindades de un píxel; y tercera, las operaciones globales que modifican la intensidad del píxel con base en alguna medida que involucra a todos los píxeles de la imagen [Alvarado Moya, 2012, Gonzáles and Wintz, 1996, Gonzales C. and Woods E., 2008].

Dentro de las operaciones puntuales, se encuentran las transformaciones exponencial y logarítmica, que sirven para ajustar el brillo de una imagen. Los operadores morfológicos son considerados operaciones locales, los cuales son útiles en tareas de segmentación. Algunas operaciones locales como el filtro promedio y el filtro de mediana han sido utilizadas para disminuir los efectos negativos que se pueden presentar en una imagen digital, en otras palabras, el ruido. Otra operación local, una de las más utilizadas, es la detección de bordes, la cual utiliza información del gradiente

para detectar cambios en la intensidad de la imagen. En este sentido, la técnica más usada es el algoritmo propuesto por John Francis Canny en 1986 [Canny, 1986], el cual ha demostrado ser bastante robusto para la detección de bordes.

Dentro de las operaciones globales se encuentra la técnica de Ecualización Adaptativa de Histograma Limitada por Contraste (CLAHE, por sus siglás en inglés) [Pizer et al., 1987]. Este método opera limitando el realce de contraste en pequeñas regiones de la imagen llamadas mosaicos, lo que mejora la calidad de la imagen [Sahu et al., 2019]. CLAHE es un método que mejora la Ecualización de Histograma (HE) y la Ecualización de Histograma Adaptativa (AHE). CLAHE básicamente se basa en la división de la imagen en varias regiones no superpuestas de tamaños casi iguales, limitando el realce de contraste que usualmente es realizado por la HE ordinaria, lo cual resulta en el realce de ruido también. Básicamente, la mejora del contraste se puede definir como la pendiente de la función que relaciona el valor de la intensidad de la imagen de entrada con la intensidad de la imagen resultante deseada [Sahu et al., 2019, Rodrıguez, 2017, Yadav et al., 2014]. Además, la mejora del contraste está directamente relacionada con la altura del histograma en determinado valor de intensidad. Por lo tanto, limitar la pendiente y recortar la altura del histograma son las mismas funciones que controlan la mejora del contraste [Reza, 2004].

Otro preprocesamiento que es recurrente en el reconocimiento de emociones en expresiones faciales es la identificación y extracción de rostros [Dino and Abdulrazzaq, 2019]. Uno de los algoritmos más populares y eficientes para la detección de rostros es el algoritmo conocido como Viola-Jones [Viola and Jones, 2004].

De manera general, el algoritmo Viola-Jones se compone de cuatro partes:

1. Selección de características: Dichas características se extraen con ayuda de ventanas Haar y consisten en una diferencia de luminancia entre bloques de píxeles.

2. Creación de la imagen integral: Esta representación permite hacer la diferencia de luminancia más eficiente, ya que se construye a partir de los valores de luminancia de la imagen.

3. Entrenamiento: El entrenamiento se hace con el algoritmo Adaboost y consiste en presentar imágenes de rostros para que se puedan identificar las características que los componen.

4. Clasificador en cascada: Por cada característica encontrada se implementa un clasificador, para que cada nueva entrada tenga que ser validada por todos los clasificadores. Si la entrada es aprobada por todos los clasificadores, esta se

identifica como rostro, de lo contrario es desechada.

2.2.2. Extracción de características

Dentro de la las técnicas tradicionales de extracción de características para el reconocimiento de emociones existen dos grupos de técnicas principales: las basadas en características geométricas y las basadas en la apariencia [Konar and Chakraborty, 2015]. Por otra parte, recientemente se ha popularizado también el uso de Aprendizaje Computacional (*ML*, por sus siglas en inglés) para este fin.

Aprendizaje Computacional: Los algoritmos de ML son conocidos por aprender la relación subyacente en los datos y con base en estas relaciones toman decisiones sin necesidad de instrucciones explícitas [Carleo et al., 2019]. Las CNNs son de los mejores algoritmos de ML para el análisis de imágenes y han demostrado un buen rendimiento en tareas de segmentación, clasificación, detección y recuperación de imágenes [Khan et al., 2020].

Características Geométricas: Este tipo de características se basan en los rasgos faciales distintivos como los ojos, la boca y la nariz que son las principales Regiones de Interés (*ROI*, por sus siglas en inglés) en el reconocimiento de expresiones faciales. En este enfoque, los investigadores se centran en representar las relaciones geométricas de la expresión facial. Un ejemplo de este tipo de técnicas son los modelos de contorno activo, los cuales pueden usarse para seguir el movimiento de los labios [Kass et al., 1988].

Enfoque Basado en Apariencia: Estas técnicas consideran al rostro como un arreglo de valores de intensidad factible para ser procesados. Normalmente, implica un preprocesamiento seguido de una codificación compacta a través de la reducción de redundancia estadística. Dentro de este grupo de técnicas, se encuentra el flujo óptico, ondeletas de Gabor y de apariencia basada en píxeles [Lowhur and Chuah, 2015, Reddy et al., 2019].

Dentro de los métodos basados en apariencia, hay métodos que buscan identificar movimiento en imágenes, uno de los más estudiados ha sido el flujo óptico, el cual refleja los cambios de la imagen durante un intervalo de tiempo [Singh and Singh, 2011]. Este representa al movimiento aparente de los patrones de brillo en una secuencia de imágenes. Con ésto, el análisis del movimiento a partir de secuencias de imágenes busca extraer parámetros que caractericen el desplazamiento de los objetos. Dicha estimación de movimiento está relacionada con los cambios temporales y espaciales de cada píxel [Díaz Salcedo and Higuera Martínez, 2006]. Este método fue

Figura 2.4: Cálculo del flujo óptico, por el método de Singh [Catrillon et al., 2008]

propuesto por [Horn and Schunck, 1981] bajo la suposición de que un punto que se mueve no cambiara significativamente de posición en un intervalo de tiempo pequeño. En la Figura 2.4, se aprecia el cálculo de flujo óptico a partir de una secuencia de imágenes extraída de la base de datos *CMU pose* [Sim et al., 2002], dónde se observa la emoción de ira.

De acuerdo con [Szeliski, 2010], el cálculo de flujo óptico es el método más general para la estimación de movimiento. Este tipo de método evalúa el movimiento de cada píxel, el cual generalmente es conocido como *óptica*. Estos métodos implican minimizar la diferencia de brillo o diferencia de color entre los píxeles correspondientes sumados en la imagen. La siguiente ecuación es una forma general para describir el cálculo del flujo óptico, ya que puede verse como la suma del cuadrado de diferencias (error) entre imágenes.

$$E_{SSD-OF}(I_i) = \sum_i [I_1(X_i + u_i) - I_0(Xi)]^2 \tag{2.1}$$

donde I_1 e I_0 son dos imágenes continuas, I_1 es la imagen actual e I_0 es la imagen anterior. X_i representa una posición determinada en la imagen y u_i el desplazamiento.

A $I_1(X_i + u_i) - I_0(X_i)$ se le conoce como error residual e_i o diferencia de marco desplazado.

Existen dos enfoques principales para aplicar este método:

- El primer enfoque consiste en realizar las operaciones sobre regiones superpuestas (ventanas).
- El segundo enfoque consiste en añadir términos de suavidad para buscar mínimos globales.

De igual forma, el manejo de la información para la estimación del movimiento se puede usar de dos formas:

- Combinando información local y global de los píxeles de una imagen.
- Realizando una combinación de modelos de movimiento locales y globales.

Una de las desventajas de este método es que el brillo no es una medida apropiada para medir consistencia (en apariencia), por lo que los cambios de iluminación entre imágenes influyen fuertemente en estos métodos, ya que afectan directamente la premisa de suposición de flujo constante, es decir, que dos imágenes consecutivas no cambian en gran manera. Una forma de evitar este problema es con el cálculo de la *fase de filtros orientados*, ya que esta medida es más robusta a cambios de iluminación.

Otro de los inconvenientes que tiene el cálculo de flujo óptico es el costo computacional debido al gran espacio de búsqueda bidimensional que se requiere para la estimación del flujo. Para solucionar este problema, la mayoría de los algoritmos utilizan variaciones del descenso de gradiente y métodos continuos *coarse-to-fine* para minimizar la función de energía global del cálculo, es decir, disminuir el número de operaciones que se ocupan para realizar el cálculo.

De igual forma, se ha utilizado una técnica llamada *combinación de movimientos*. La idea básica de esta técnica es sustituir porciones de la estimación actual con hipótesis generadas por técnicas más básicas y alternarlas con el descenso de gradiente local para minimizar el cómputo.

Las aplicaciones principales de este tipo de métodos son tres:

- Estimación de movimiento en varios cuadros.
- Eliminación de ruido en videos.
- Desentrelazado de videos.

2.2.3. Clasificadores para reconocimiento de emociones a través de expresiones faciales

Usualmente, en el análisis de expresiones faciales se busca identificar las emociones de sorpresa, ira, felicidad, miedo, neutral, asco y tristeza. Para realizar esta tarea,

existe el enfoque tradicional que incluye al método de los K-Vecinos Más Cercanos, Máquinas de Soporte Vectorial, Bosques Aleatorios y Árboles de Decisión [Dino and Abdulrazzaq, 2019, Konar and Chakraborty, 2015, Dino et al., 2020]. Otro enfoque más reciente hace uso de las redes neuronales profundas, siendo las CNNs como el método de más frecuente uso [Moolchandani et al., 2021, Anvita et al., 2020].

De manera más específica, las Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés) son una de las técnicas más rápidas y sencillas utilizadas en el reconocimiento de emociones. Este método pertenece a los algoritmos de aprendizaje supervisado. De manera general, SVM se utiliza para regresión y clasificación [Anvita et al., 2020].

Por otro lado, las CNNs han demostrado tener un mejor desempeño en el reconocimiento de emociones mediante las expresiones faciales [Anvita et al., 2020]. Este tipo de redes requiere siempre un gran número de imágenes para obtener buenos resultados de clasificación. Una práctica que se realiza es aumentar el conjunto de datos con diferentes cambios para generar nuevas imágenes con pequeños cambios en las apariencias y posturas. El principal beneficio de CNN es que el número de parámetros es menor en comparación con otras redes, por lo cual su entrenamiento es más rápido.

2.3. Reconocimiento de emociones mediante voz

El reconocimiento de emociones mediante la voz es un campo importante en el procesamiento del lenguaje. De manera general, el reconocimiento de emociones mediante la voz sigue la metodología de la Fig. 2.5 [Konar and Chakraborty, 2015, Issa et al., 2020], en la cual se destacan tres módulos principales: preprocesamiento, extracción de características y clasificación; la extracción de características es un tema de especial interés para los investigadores.

2.3.1. Preprocesamiento de voz

El término preprocesamiento se refiere a todas las operaciones que se deben realizar en las muestras de la señal de voz antes de extraer las características. Este preprocesamiento es necesario debido a que las muestras generalmente contienen información no deseada como el ruido sensado del ambiente durante la grabación. El ruido en el audio puede eliminarse mediante el preprocesamiento basado en filtros

[Koduru et al., 2020]. Otra técnica de preprocesamiento usada recientemente es la transformación de la información de audio a imagen, a través de una transformación en el dominio de frecuencia. Lo anterior es una representación de la distribución de energía del sonido en función de sus componentes en frecuencia. En otras palabras, el espectro representa el nivel de presión sonora en función de la frecuencia. [Hossain and Muhammad, 2019, Ramírez Cornejo and Pedrini, 2019, Fan et al., 2016].

2.3.2. Extracción de características en voz

Dentro del análisis de voz, existen diversas características que pueden agruparse de diferente manera. Dentro de estas agrupaciones, se identifican dos grupos importantes [Pérez Espinosa and Reyes García, 2010]: el análisis temporal y el análisis espectral.

El análisis temporal se enfoca en analizar la voz sin hacer una transformación al dominio de frecuencia. Dentro de este análisis se encuentran diferentes tipos de ca-

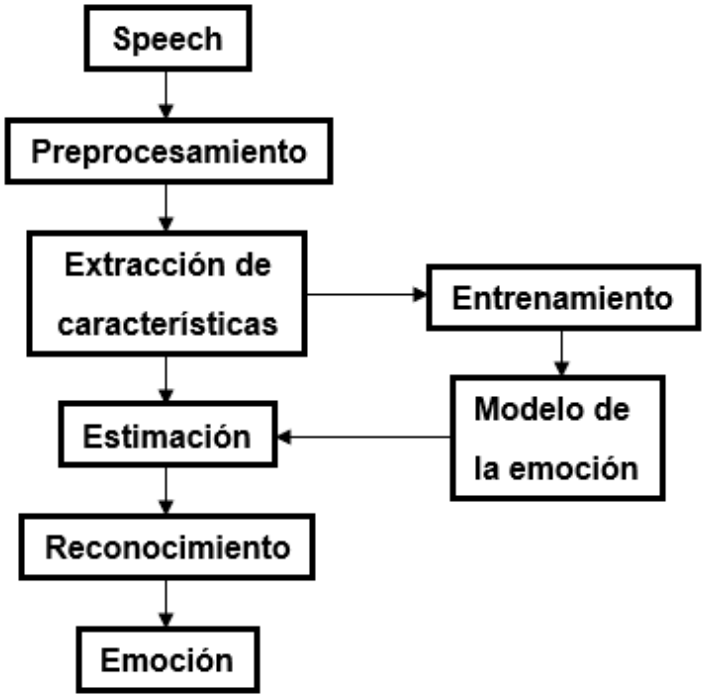


Figura 2.5: Metodología general para el reconocimiento de emociones a partir del análisis de voz [Konar and Chakraborty, 2015]

racterísticas, como por ejemplo las características prosódicas, continuas y de calidad de voz.

Características prosódicas: Este tipo de característica refleja una actitud o estado emocional del emisor [Pérez Espinosa and Reyes García, 2010]. También conlleva a atributos extralingüísticos que aportan información sobre las características del locutor, como su edad, sexo y cultura. Lo anterior se debe a que el humano mientras habla, produce patrones específicos de duración, entonación e intensidad. La prosodia puede considerarse como una característica del habla asociada a unidades como las sílabas, palabras, frases y oraciones. En consecuencia, la prosodia suele considerarse como información suprasegmental (de elementos fonéticos) [Koolagudi and Rao, 2020].

Características continuas: Las características continuas del habla se han utilizado mucho en el reconocimiento de emociones. Este tipo de características están relacionadas con la frecuencia fundamental, la energía, la tasa de articulación y la información espectral en las partes sonoras y no sonoras. A partir de la frecuencia fundamental y de la energía se puede obtener la media, desviación estándar, coeficientes de regresión lineal y otras características de cuarto orden. De la duración se puede extraer el ritmo del habla, la relación entre la duración de las regiones con y sin voz y la duración del discurso vocal [Semwal et al., 2017].

Características de calidad de voz: Hay estudios experimentales donde se muestra que hay una relación estrecha entre la calidad de la voz y la emoción percibida. Una amplia gama de variables fonéticas contribuye a la impresión subjetiva de la calidad de la voz. Estas variables pueden clasificarse en: nivel de voz (donde entra la amplitud, la energía y la duración de la señal), tono de voz, estructura temporal, límites de frases, fonemas, palabras y rasgos de la voz [Ayadi et al., 2011].

Por otra parte, el análisis espectral se enfoca en la representación del espectro de audio a partir de determinadas transformaciones. Lo interesante de estas representaciones es que, pueden extraerse atributos de manera más sencilla en comparación con una representación en el dominio temporal. Dentro de las representaciones espectrales habituales se encuentran espectrogramas, cromagramas y cocleogramas.

Cromagrama: Es una representación visual de la energía distribuida en las 12 clases de tono que equivalen a las 12 notas musicales. Este tipo de representación espectral normalmente es usado en la identificación automática de instrumentos. En la Figura 2.6, se representa la escala cromática en la que el eje chroma representa cada una de las notas musicales. [Birajdar and Patil, 2020]

Espectrogramas: El espectrograma es una representación visual de las dife-

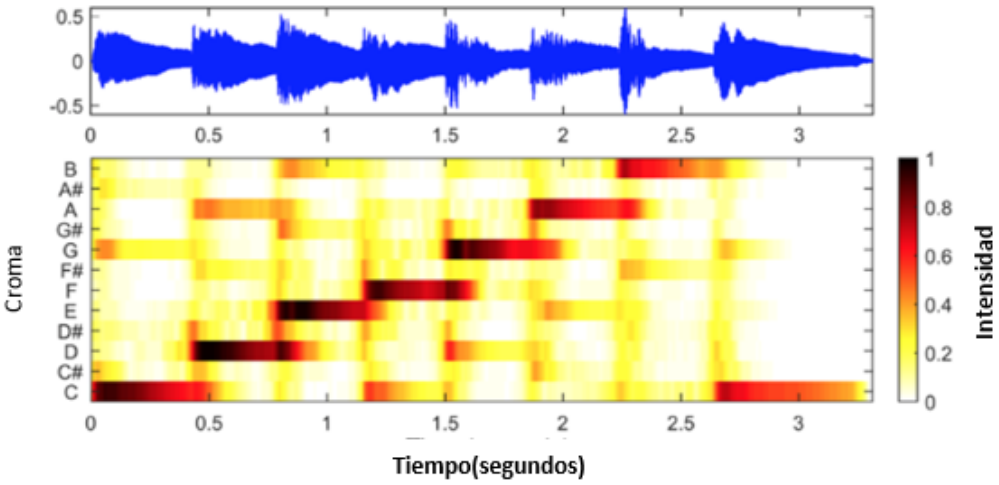
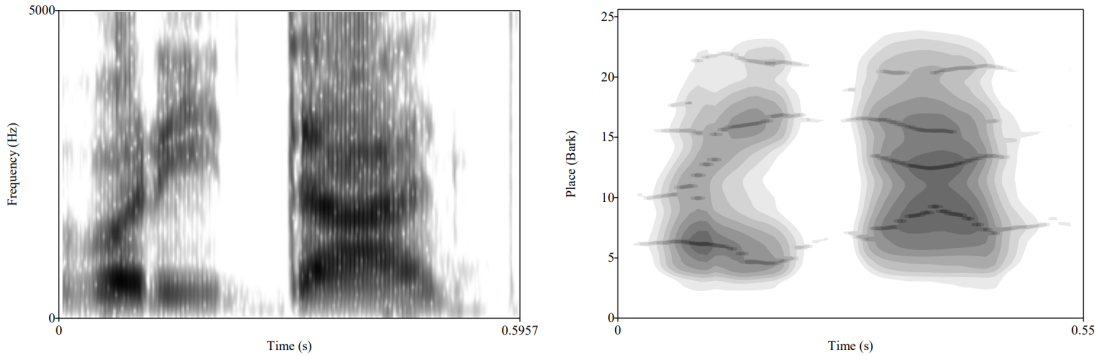


Figura 2.6: Ejemplo de un cromagrama

rentes frecuencias en el eje vertical, así como de la intensidad del sonido mediante diferentes colores a lo largo del tiempo que se representa en el eje horizontal. En el reconocimiento de emociones, la información más valiosa está contenida en la forma espectral del tracto vocal; esto es, debido a la naturaleza variable de la voz [Chandrasekar et al., 2014]. Algunas de las representaciones más utilizadas en los espectrogramas son la Transformada de Fourier (FT) y los Coeficientes Cepstrales de Frecuencia de Mel (del inglés, MFCC) [Meng et al., 2019, Zhao et al., 2019].



(a) Espectrograma basado en la transformada de Fourier

(b) Cocleograma equivalente

Figura 2.7: Comparación entre un espectrograma y un cocleograma [Cicres, 2009]

Cocleogramas: Es una representación que emula la excitación de los filamentos del nervio auditivo de la membrana basilar de los seres humanos; esta se encuentra en la cóclea del oído interno. Este tipo de representación refleja la riqueza en el análisis fonético-acústico del habla. Las principales características de esta representación son

[Cicres, 2009, Peng et al., 2021]:

- La unidad temporal se mide en segundos.
- La unidad de frecuencia se mide en Barks. La escala psicoacústica tiene un rango de 1 a 24 Barks y corresponde a las primeras 24 bandas críticas del oído.
- Se analiza la excitación de la membrana basilar por unidad de tiempo.

En la Figura 2.7, se contrastan dos representaciones espectrales equivalentes donde se observa que el cocleograma muestra una distribución más clara de la intensidad del sonido a través del eje Y (región baja y media), así como el desvanecimiento de la señal en zonas de baja intensidad.

2.3.3. Clasificadores para reconocimiento de emociones a través de la voz

En la literatura, de manera tradicional, se han explorado distintos clasificadores de patrones para el desarrollo de sistemas de reconocimiento de emociones. Estos métodos se pueden clasificar en lineales y no lineales. Los clasificadores lineales realizan la separación tomando una decisión basada en el valor de una combinación lineal de las características del objeto. Estas características también se conocen como valores de características y suelen presentarse al clasificador en forma de un arreglo denominado vector de características. En cambio, los clasificadores no lineales utilizan combinaciones ponderadas de características no lineales [Koolagudi and Rao, 2020, Chandrasekar et al., 2014, Ayadi et al., 2011].

Los modelos de clasificación tradicionales más populares en el reconocimiento de emociones mediante voz son los Modelos Ocultos de Markov, Modelos de Mezclas Gaussianas, Máquinas de Soporte Vectorial, Clasificador Naïve Bayes y K-Vecinos Más Cercanos [Reddy and Kuchibhotla, 2019].

2.4. Métodos para el reconocimiento de emociones utilizando información audiovisual

Como se ha mencionado en el Capitulo 1, el reconocimiento de emociones usando información audiovisual ha incrementado el interés de investigación, ya que las

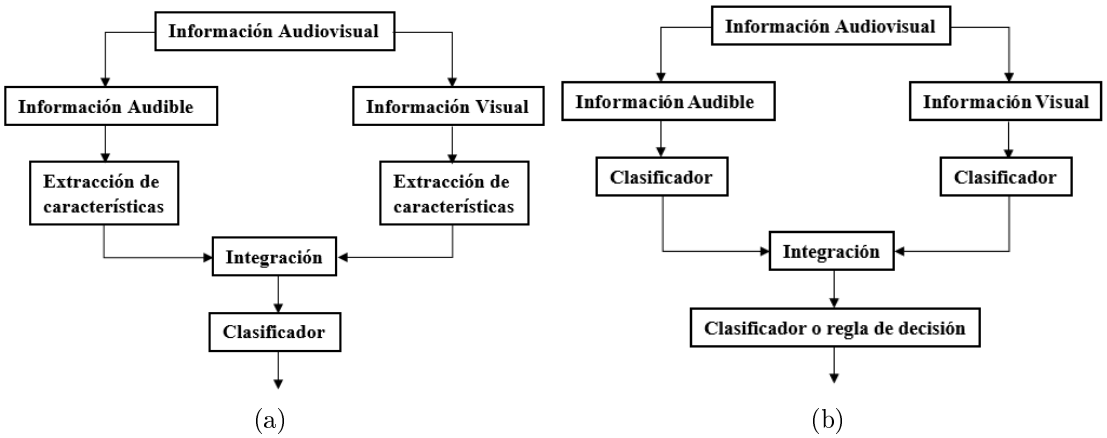


Figura 2.8: Dos metodologías para el reconocimiento de emociones a través de información audiovisual [Wu et al., 2014, Avots et al., 2019]. a) Integración de características y b) Integración de decisión

personas expresan sus emociones a través de diferentes modalidades. Al tomar en consideración más fuentes de información, los enfoques multimodales permiten una mejor estimación de las emociones humanas por lo que aumenta la fiabilidad de los resultados y disminuye el nivel de ambigüedad con respecto a las emociones entre los distintos canales de comunicación [Konar and Chakraborty, 2015, Wu et al., 2014, Avots et al., 2019]. El uso de los canales de comunicación verbal y no verbal permite crear un sistema en el que el estado emocional se expresa con mayor claridad [Avots et al., 2019].

De manera general, existen dos metodologías para realizar el reconocimiento de emociones a través de información audiovisual [Wu et al., 2014, Avots et al., 2019]. El primer enfoque se realiza combinando las características auditivas y visuales extraídas en un único vector de características, que luego pasa a ser modelado por un único clasificador y por tanto, realiza el reconocimiento de emociones [Trigeorgis et al., 2017, Hossain and Muhammad, 2019]. En el segundo enfoque, las características extraídas de las diferentes modalidades se procesan de forma independiente y los resultados de clasificación individual o por modalidad se combinan en un solo nivel de decisión [Hossain and Muhammad, 2019]. En la Figura 2.8 se muestran de manera gráfica ambas metodologías.

2.4.1. Enfoque basado en la integración de características

[Ramírez Cornejo and Pedrini, 2019] muestra un ejemplo de este enfoque que consiste en extraer las expresiones faciales de muestras de video y dividirlas en dos imágenes que ubican diferentes regiones de interés. La primera región se ubica en los ojos y la segunda, en la boca, a estas se les aplica una ecualización CLAHE y una transformada Census. A partir de estas imágenes, se extraen características con CNNs. Las características obtenidas a partir de imagen se fusionan, posteriormente, con las características obtenidas a partir de audio. Las características del audio son extraídas de igual forma con CNNs a través de una representación espectral del audio, en específico el Espectrograma Logarítmico de Mel. El vector resultante de ambas características fusionadas se reducen con las técnicas Análisis de Componentes Principales y Análisis Discriminante Lineal (PCA y LDA respectivamente, por sus siglas en inglés). Para la parte de clasificación se utilizan clasificadores tales como SVM, K-NN y Regresión Lineal.

Otro ejemplo es el mostrado en [Hossain and Muhammad, 2019]. De igual forma, se ocupan CNNs como extractores de características, solo que éstas son combinadas utilizando aprendizaje computacional extremo y finalmente clasificadas por una SVM. La información del audio es preprocesada mediante técnicas de ventanas y encuadre, para posteriormente obtener su representación espectral, mediante un espectrograma de Mel y así poder utilizar la CNN como extractor de características. Por otro lado, las imágenes de expresiones faciales se trabajaron en escala de grises y únicamente se aplicó una normalización como preprocesamiento.

En el trabajo de [Trigeorgis et al., 2017], el audio se trabajó en su representación temporal, extrayendo características acústicas con ayuda de redes LSTM. En el caso de las expresiones faciales, las imágenes se procesaron mediante la Transformación de Características Invariante al Escalado (*SIFT*, por sus siglas en inglés) y su Histograma de Gradientes Orientados (HOG, por sus siglas en inglés). La características extraídas de audio e imagen son unidas en un solo vector y clasificadas por dos capas de redes LSTM. A diferencia de otros trabajos, la salida del método propuesto pueden tener dos valores que corresponden al *aerousal* y la *valencia*, los cuales representan otro modelo de emociones diferente al de [Ekman et al., 1983]. Los valores de aerousal y valencia están representados en dos ejes: el eje aerousal representa que tan positiva o negativa es una emoción y el eje valencia representa qué tan intensa es la emoción.

2.4.2. Enfoque basado en la integración de decisión

La aportación principal de este trabajo [Fan et al., 2016] consiste en la implementación de un clasificador que combina CNNs con RNNs, ya que estas últimas son capaces de detectar dependencias temporales. Este clasificador se aplicó a una secuencia de imágenes que fueron extraídas de una muestra de video, con el fin de identificar la emoción representada en el video. Por otro lado, las características que son extraídas del audio se clasifican con una SVM. Finalmente, la fusión de la decisión de ambos clasificadores está dada por una suma, es decir $W_{audio} + W_{imagen}$, donde cada factor W es un vector de probabilidad de decisión el cual es utilizado para seleccionar a la emoción con mayor probabilidad.

Otro ejemplo de este enfoque es el presentado por [Wang et al., 2020]. En este trabajo, la información de audio es clasificada mediante el uso de su representación espectral por una red LSTM-CNN, ocupando la LSTM como extractor de características. Por otro lado, la imagen es clasificada por una CNN-RNN, en la que la CNN se ocupa como extractor de características y la RNN como clasificador. La combinación de decisión está dada por la ecuación $argmax(W_0 S_{face} + W_1 S_{speech})$, donde W_0 y W_1 son dos pesos de decisión, cuya suma es 1 y S_{face} y S_{speech} representan los vectores de salida en cada clasificador.

Otro trabajo que ocupa este enfoque es el presentado por [Antoniadis et al., 2021], el cual trabaja bajo las premisas de poca iluminación, baja resolución y vista de diferentes ángulos. En este caso, tanto la información de audio como de imagen se clasifican con las arquitecturas ResNet-50 y LSTM, ocupando la ResNet-50 como extractor de características y la LSTM como clasificador. En el caso de la imagen se entrenan tres redes diferentes, una para el rostro, otra para el cuerpo y una última para el contexto; mientras que el audio se procesa sólo con su representación espectral. La combinación de decisión se hace con una función probabilística.

2.5. Aprendizaje profundo

El aprendizaje profundo es un subcampo de las ciencias de la computación y una rama de la inteligencia artificial. El objetivo principal de este aprendizaje es reconocer patrones en la información a partir de la experiencia (ejemplos), y parte de su éxito se debe al hecho de la mejora significativa de las tecnologías existentes (por ejemplo, el reconocimiento de objetos en imágenes) [Carleo et al., 2019]. Estos avances constituyen manifestaciones del impacto que los métodos de ML pueden tener en tareas especializadas [Carleo et al., 2019]. Particularmente, el uso de Redes

Neuronales Artificiales (ANNs, por sus siglas en inglés) es un tema bastante usado, por lo cual vale la pena profundizar en esta área.

2.5.1. Introducción

Las ANNs son una herramienta que imita de manera muy básica el funcionamiento de las neuronas biológicas. Estas redes están organizadas en una estructura compleja de interconexiones llamadas capas, las cuales poseen conexiones que permiten la interacción entre ellas. Además, se les conoce como un sistema de computación constituido por un gran número de elementos simples de procesamiento interconectados, que procesan información en respuesta a algún estímulo externo [Haykin, 2009].

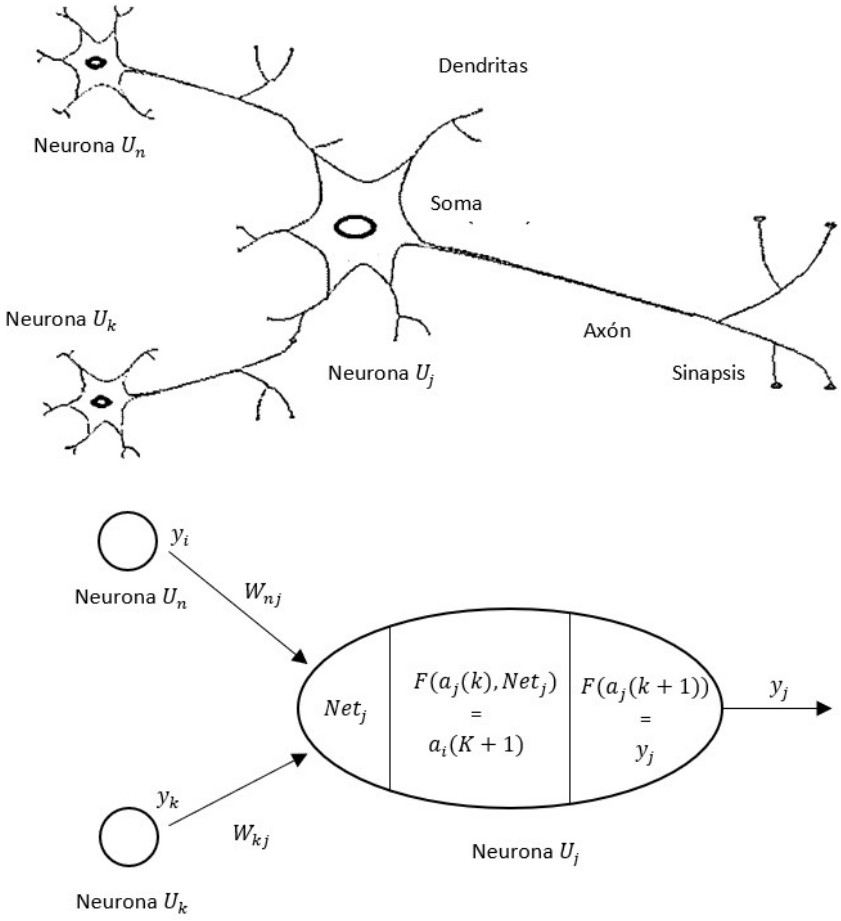


Figura 2.9: Esquema de una neurona biológica comparado con el esquema de una neurona artificial [Pajares Martinsanz et al., 2010]

Una de las principales características de las ANNs es que se conciben como un esquema computacional distribuido, que básicamente se asimila a una estructura del sistema nervioso de los seres humanos [Hilera Gonzales and Martínez Hernando, 1995].

La unidad básica de procesamiento de las redes neuronales artificiales es la neurona artificial, la cual conserva de la neurona biológica el procesamiento del núcleo, las dendritas (entradas) y el axón (salida). La Figura 2.9 muestra los componentes de una neurona artificial y una neurona biológica.

Uno de los primeros modelos exitosos de neurona artificial es el perceptrón, el cual fue propuesto por [Rosenblatt, 1958], donde se destacan los siguientes elementos (ver Figura 2.10):

- Los valores de entrada X_1, X_2, ... X_m son los estímulos que recibe la neurona, los cuales pueden provenir de otra neurona o perceptrón.
- Enlaces de conexión, los cuales son parametrizados por los pesos sinápticos W_1, $W_2 ... W_m$.
- El punto de suma se encarga de agregar las diferentes señales ponderadas que llegan a la neurona. Cabe señalar que, cada señal tiene un peso W_i diferente.
- La función de activación define la manera en que una neurona se activa, de

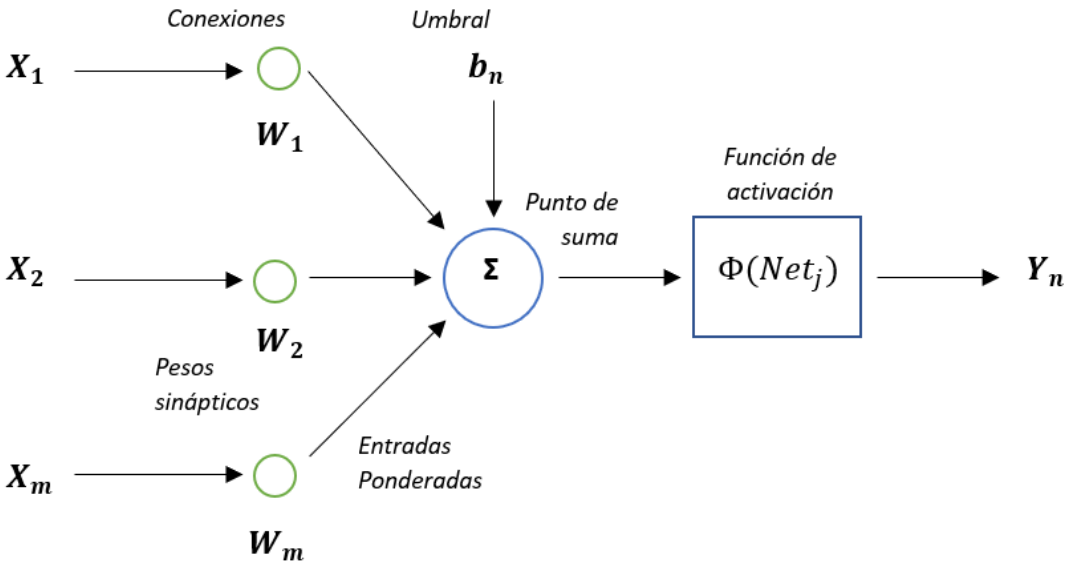


Figura 2.10: Esquema de un perceptrón [Nolasco Martínez et al., 2013]

acuerdo con una función del potencial de activación. Existen varios tipos de funciones tales como: la función escalón o umbral, función lineal y función sigmoidal, entre otras.

- La señal de salida Y_n puede interconectarse con otras neuronas.

Dentro de las funciones de activación mas comunes tenemos (ver Figura 2.11):

- **Identidad:** Esta función, como su nombre lo indica, no hace ninguna transformación de los valores que recibe, ya que la entrada será igual a la salida. También es llamada función lineal, debido a su comportamiento.

- **Escalón:** Para todos los valores menores o iguales a 0 esta función arrojará el valor cero, de lo contrario 1; hay una variación en la cual los valores que arroja son -1 para valores menores o iguales a 0 y 1 para el caso contrario.

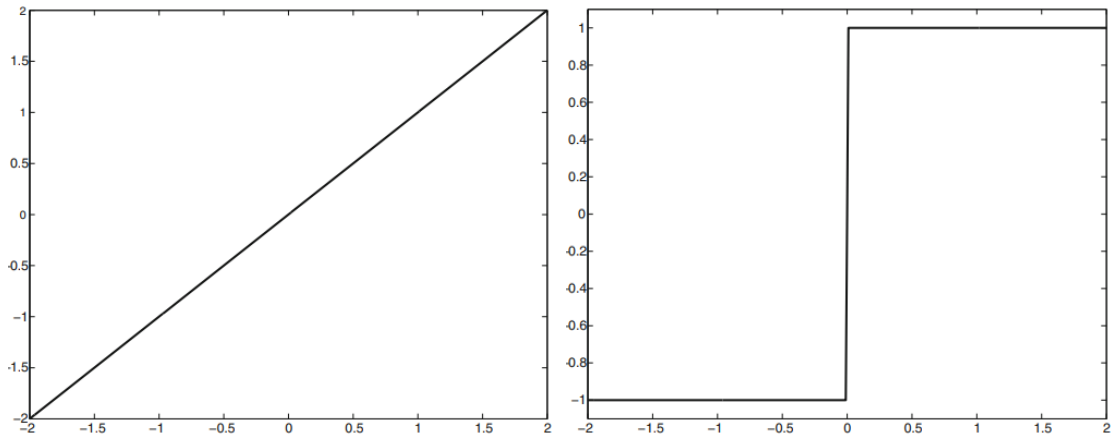
- **Sigmoidal:** Los valores de salida de esta función están acotados entre los valores 0 y 1, por lo que la respuesta de esta función podría interpretarse como una probabilidad. La función es la siguiente:

$$f(x) = \frac{1}{1 - e^{-x}} \tag{2.2}$$

- **ReLU:** Para valores negativos esta función arroja el valor 0 y para valores positivos esta función se comporta como la función identidad.

- **SoftMax:** La función Softmax transforma las salidas a una representación en forma de probabilidades, de tal manera que el sumatorio de todas las probabilidades de las salidas es de 1.

Una de las arquitecturas de redes neuronales más populares en la actualidad es el perceptrón multicapa [Haykin, 2009], también conocido como redes de propagación hacia el frente (FFNN, por sus siglas en inglés). Esta arquitectura está constituida por varios perceptrones organizados en forma de capas e interconectados entre sí; su objetivo principal es estimular la red desde sus entradas, propagando la información de una capa a otra, hasta finalmente obtener una salida. Cabe señalar que existe una distinción entre redes neuronales profundas y no profundas, siendo la principal diferencia entre estas el número de capas ocultas. Es decir, una red neuronal poco profunda consta de pocas capas, mientras que las redes profundas constan de muchas capas ocultas [Goodfellow et al., 2016].



(a) Función Lineal

(b) Función Escalón

(c) Función Sigmoidal

(d) Función ReLU

Figura 2.11: Representación gráfica de diferentes funciones de activación [Aggarwal C., 2018]

Principales arquitecturas de redes neuronales profundas

Existen diferentes arquitecturas profundas que se han propuesto a lo largo de los años (Fig. 2.12) para dar solución a diferentes problemas. Para el análisis de imagen, el modelo más usado es la CNN, previamente ya señalada [Aggarwal C., 2018]. A continuación se presentan diferentes modelos de red y sus aplicaciones:

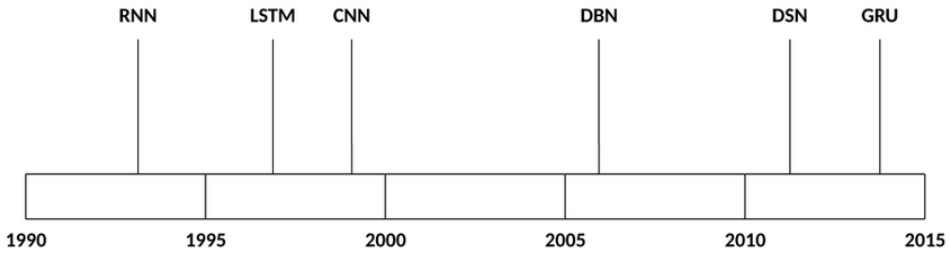


Figura 2.12: Línea del tiempo de la creación de diversas arquitecturas [Developer, 2021]

Redes Neuronales de Propagación Hacia el Frente

El principal atractivo de la red FFNN es su capacidad de aproximación universal. Una de las aplicaciones de las FFNN es el reconocimiento facial [Yang and Ma, 2019]. Generalmente, una FFNN consta de una capa de entrada, un conjunto de capas ocultas y una capa de salida (ver Figura 2.13). La salida de este tipo de red está en función de los valores de entrada y del conjunto de pesos sinápticos que se encuentran entre un nodo y otro.

El objetivo de este tipo de red es obtener de forma iterativa un vector gradiente, en el que cada componente se define como la derivada del error con respecto a un parámetro; ésto se hace mediante la regla de la cadena. Al procedimiento de encontrar el vector gradiente, que tiene como finalidad encontrar el incremento que deben sufrir dichos pesos en una estructura de red, se le conoce como *algoritmo de retropropagación*; recibe este nombre debido a que para estimar el mejor conjunto de pesos de la red, es necesario que el error se propague desde la salida de la red hacia sus entradas, calculando entre cada capa de forma parcial el vector gradiente del error. Una vez que se obtiene el gradiente, se pueden aplicar varias técnicas de optimización derivativas para actualizar los pesos de la red. Inclusive, se pueden aplicar técnicas de optimización heurísticas, tales como los algoritmos genéticos [Roger Jang, 1997], las cuales no requieren del cálculo del gradiente. El paradigma de aprendizaje resultante, a menudo, se denomina regla de aprendizaje de retropropagación.

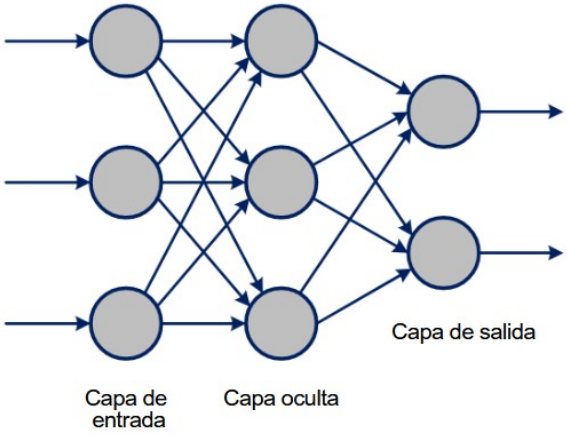


Figura 2.13: Esquema de un perceptrón multicapa [Roger Jang, 1997]

Redes Neuronales Recurrentes

Esta red se caracteriza por la existencia de lazos de realimentación. Estos lazos se realizan entre neuronas de diferentes capas, neuronas de la misma capa o entre una misma neurona. Esta estructura permite que esta red sea adecuada para estudiar la dinámica de sistemas no lineales, tal como se muestra en la Figura 2.14. Un inconveniente de este tipo de redes es que son difíciles de entrenar, debido a que sus capas temporales dependen de sus entradas mismas. Sin embargo, se han desarrollado algoritmos para su entrenamiento como el *Backpropagation in time* que fue diseñado para entrenar este tipo de redes y así poder trabajar con una secuencia, por ejemplo con series de tiempo [Aggarwal C., 2018].

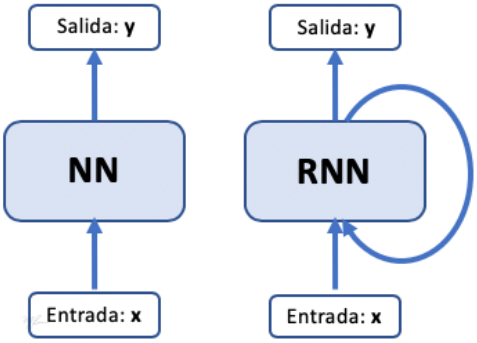


Figura 2.14: Estructura de una red neuronal tradicional comparada con una recurrente. Imagen basada en [Aggarwal C., 2018]

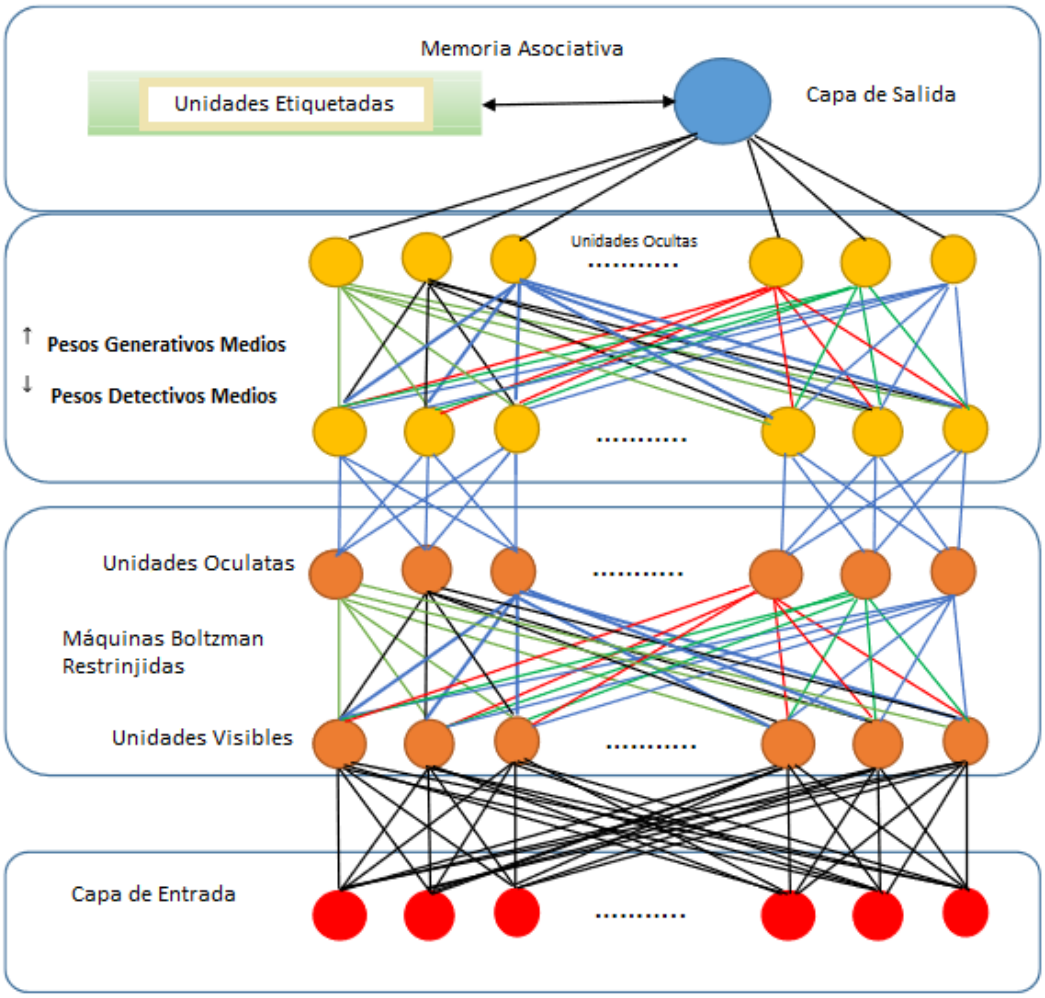


Figura 2.15: Estructura de una red DBN

Redes de Creencia Profunda

Las Redes de Creencia Profunda (DBN, por sus siglas en inglés) se inventaron como solución para los problemas que surgen al entrenar redes neuronales tradicionales, tales como el aprendizaje lento o el estancamiento debido a los mínimos locales. Las DBN constan de varias capas de neuronas (Fig 2.15), también conocidas como máquinas de Boltzmann. Las Máquinas de Boltzmann Restringidas (RBM, por sus siglas en inglés) son de carácter binario. El valor del estado se determina de acuerdo con métodos de probabilidad y estadística. Lo anterior permite que una variable binaria determine la salida de la red. A su vez, este proceso permite el entrenamiento por capa de la red que conlleva a una mejor optimización en la búsqueda de los pesos.

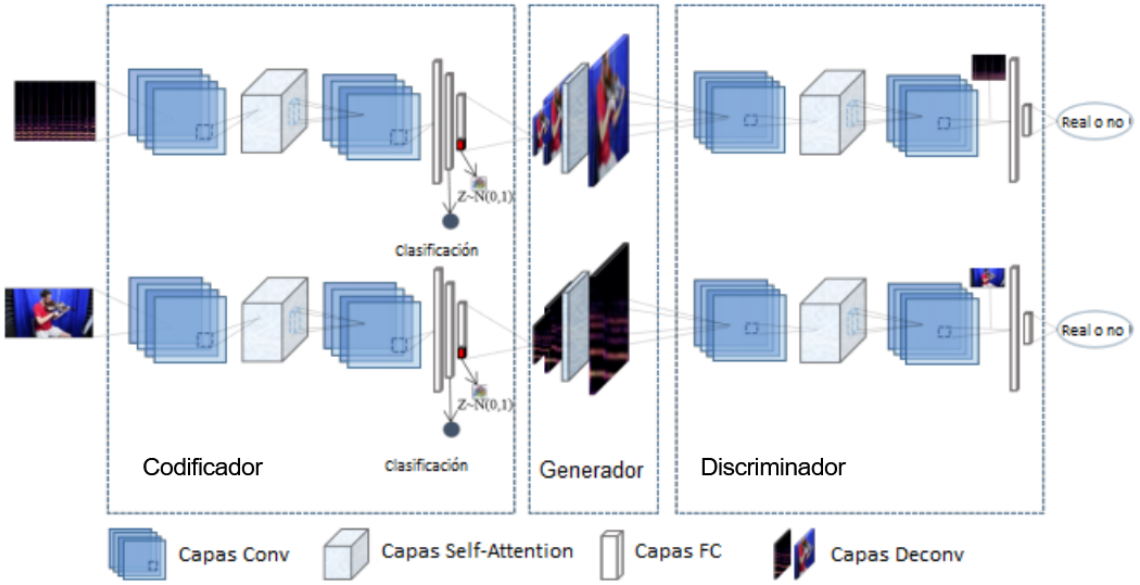


Figura 2.16: Ejemplo de un modelo GAN [Tan et al., 2020]

Redes Generativas Antagónicas

La arquitectura de Redes Generativas Antagónicas (GAN's, por sus siglas en inglés) fue desarrollada recientemente, cuya aplicación más popular es en imágenes. Esta red se caracteriza porque es capaz de generar imágenes a partir de otras. Esta red ha sido utilizada en la generación de información audiovisual, con el fin de identificar instrumentos musicales en clips de video. La configuración usada se muestra en la Figura 2.16 [Tan et al., 2020].

Redes de Apilamiento Profundo

Las Redes de Apilamiento Profundo (DSN, por sus siglas en inglés) son un tipo especial de modelo equipado con aprendizaje paralelo y escalable. Las DSN tienen ventajas sobre otros modelos por su simplicidad en el aprendizaje.

La filosofía del diseño DSN se basa en el concepto de apilamiento, donde los módulos simples de funciones o clasificadores se componen y luego se "apilan". Siguiendo esta filosofía, la arquitectura DSN consta de muchos módulos de apilamiento, cada uno de los cuales toma una forma simplificada de perceptrón multicapa, utilizando optimización convexa para aprender los pesos del perceptrón [Deng et al., 2013]. Se ocupan principalmente en recuperación de información y reconocimiento continuo de la voz.

Memoria a corto y largo plazo

Las unidades de Memoria a Corto y Largo Plazo (LSTM, por sus siglas en inglés) son un tipo de bloques empleados para la construcción de redes neuronales. Los bloques se caracterizan por ser capaces de recordar diferentes valores a lo largo de intervalos de tiempo aleatorios. Básicamente, pueden recordar estados previos y utilizar esta información para decidir cual será el siguiente estado. Esta característica las hace adecuadas para manejar series cronológicas [Graves, 2012].

Los bloques LSTM han sido utilizados en diferentes áreas como: lenguaje natural, reconocimiento de caracteres manuscritos, reconocimiento de voz, reconocimiento de gestos y captura de imágenes.

2.5.2. Redes neuronales convolucionales

Dentro de los sentidos humanos, la visión es uno de los más importantes. Históricamente, se ha buscado darle este sentido a las máquinas. Las CNNs ayudan a realizar este tipo de tareas de manera automática y en la actualidad, este es uno de los tipos de red neuronal más populares tomando su nombre debido a la operación lineal matemática entre matrices llamada *convolución*. Esta arquitectura de aprendizaje profundo se ha utilizado de manera exitosa en campos como reconocimiento facial, detección de objetos y conducción autónoma [Sajjad et al., 2020, Ji et al., 2021, Aladem and Rawashdeh, 2020]. El éxito de las CNNs en tareas de visión computacional se debe a que esta red imita el funcionamiento del córtex visual humano a través de sus diferentes capas, las cuales son [De Marchi and Mitchell, 2019, Albawi et al., 2017, Tianmei et al., 2017]:

- Capa de entrada
- Capa convolucional
- Capa de submuestreo
- Capa completamente conectada
- Capa de salida

Capa de entrada: La capa de entrada recibe una imagen, la cual es una representación espacial altamente correlacionada y puede representarse digitalmente

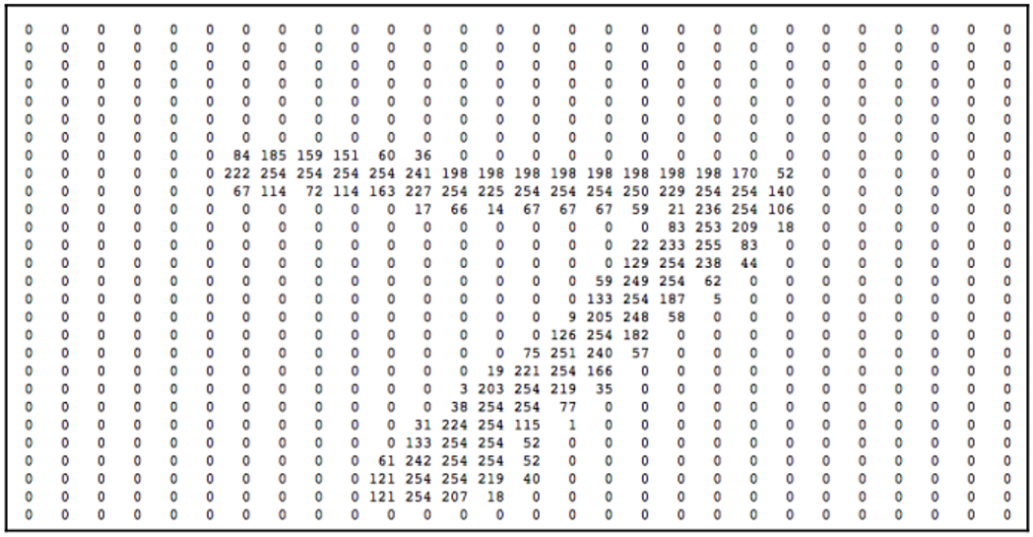


Figura 2.17: Ejemplo de la representación digital de una imagen en escala de grises [De Marchi and Mitchell, 2019]

mediante una matriz de píxeles. Dependiendo del formato de la imagen que se esté utilizando, los valores de los píxeles pueden estar dentro de diferentes rangos, por ejemplo, en una imagen en escala de grises con una codificación de 8 bits los valores asignados a cada píxel están dentro del rango de 0 a 255, donde 0 representa el color negro y 255 el blanco. En la Figura 2.17 se aprecia una representación digital de una imagen en escala de grises [De Marchi and Mitchell, 2019].

Capa convolucional: La capa convolucional es la más representativa de una CNN [Bhardwaj et al., 2018, Vasilev et al., 2019]. En esta capa se realiza la operación de *convolución*, la cual es una operación de correlación que consiste en hacer un barrido a lo largo y ancho de toda la imagen mediante una suma ponderada entre un campo receptivo (que es una parte de la imagen) y un filtro conocido como *kernel* o detector de características, normalmente con un tamaño de $n \times n$ y n siendo impar. Esta operación relaciona un píxel con sus píxeles vecinos, lo cual conlleva una correlación espacial alta. En la práctica, en cada capa convolucional se usa más de un filtro, debido a que cada filtro aprende una característica diferente. Matemáticamente, la convolución se describe con la Ecuación 2.3.

$$s(t) = (x \star w)(t) = \sum_{a=-\infty}^{\infty} x(a) * w(t-a) \tag{2.3}$$

donde x es la entrada, w es el conjunto de pesos, s es la salida, t es el tiempo y a es

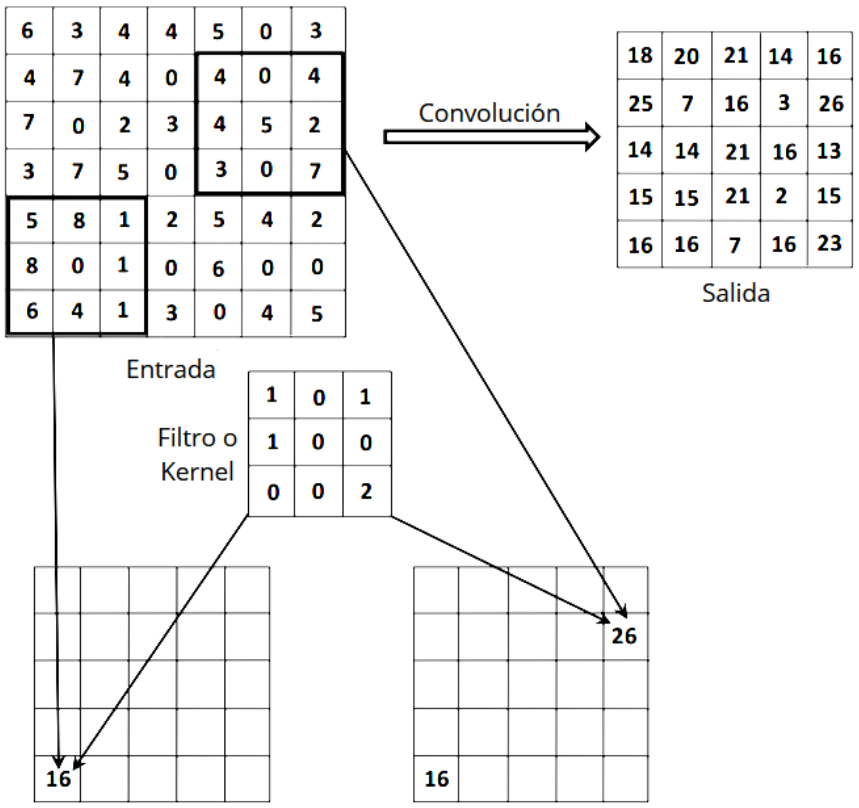


Figura 2.18: Ejemplo de una convolución entre una entrada de $7 \times 7 \times 1$ y un filtro de $3 \times 3 \times 1$ [Aggarwal C., 2018]

la época de medición.

Otra forma de ver la convolución es como una multiplicación de matrices ponderada que ocurre entre el filtro y la imagen. En la Figura 2.18 se ilustra el proceso de esta operación, en la cual puede observarse que, para cada posición posible de la entrada, se realiza un producto punto entre el filtro y la entrada hasta formar una nueva matriz; a esta matriz se le conoce como *mapa de características*. En cada capa convolucional, se pueden extraer diferentes cantidades de mapas de características. En la operación de convolución se introducen los conceptos de paso (*stride*) y relleno (*padding*), los cuales son descritos a continuación.

El *stride* o paso hace referencia al número de posiciones que avanza el *kernel* a lo largo de la imagen. Debido a que los píxeles vecinos están fuertemente correlacionados, un número grande de paso implicará una pérdida significativa de información. En la Figura 2.19 se puede ver el avance del *kernel* a lo largo de una imagen, en este caso se está usando un $paso = 2$, los pasos se miden de centro a centro.

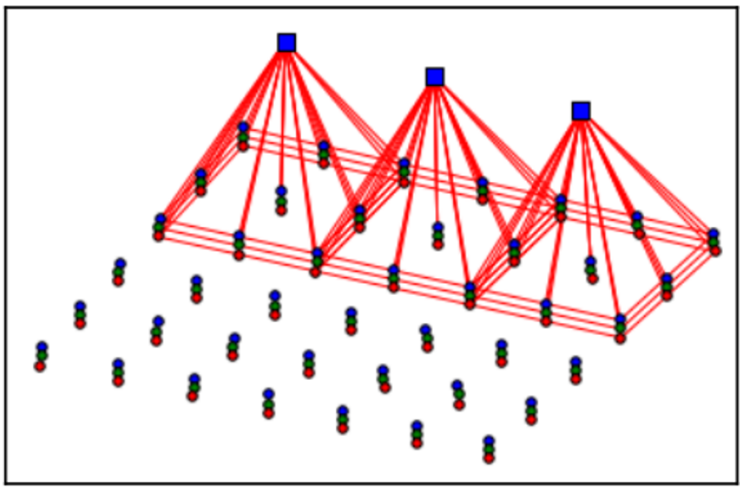


Figura 2.19: En la imagen se muestra como el filtro va avanzando dos pasos cada ocasión [Vasilev et al., 2019]

Por otro lado, el *padding* o relleno consiste en rellenar con ceros todos los bordes de una imagen (o mapa de características) antes de realizar la operación de convolución. Lo anterior, con el fin de mantener las mismas dimensiones de entrada y salida después de cada operación de convolución. En la Figura 2.20, se muestra un ejemplo de *padding*, en la que los puntos azules representan los valores originales, mientras que los puntos blancos, los ceros agregados en los bordes. Las dimensiones de salida de la operación de convolución se puede calcular con la ecuación 2.4.

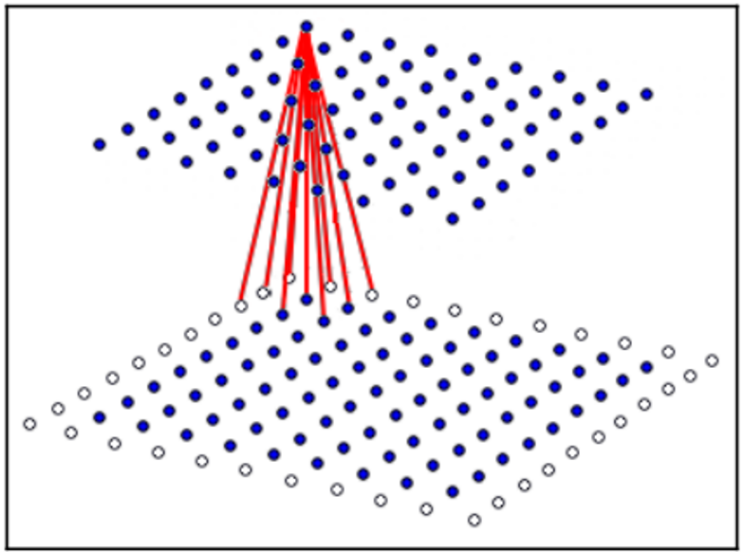


Figura 2.20: Capa convolucional con $padding = 1$ [Vasilev et al., 2019].

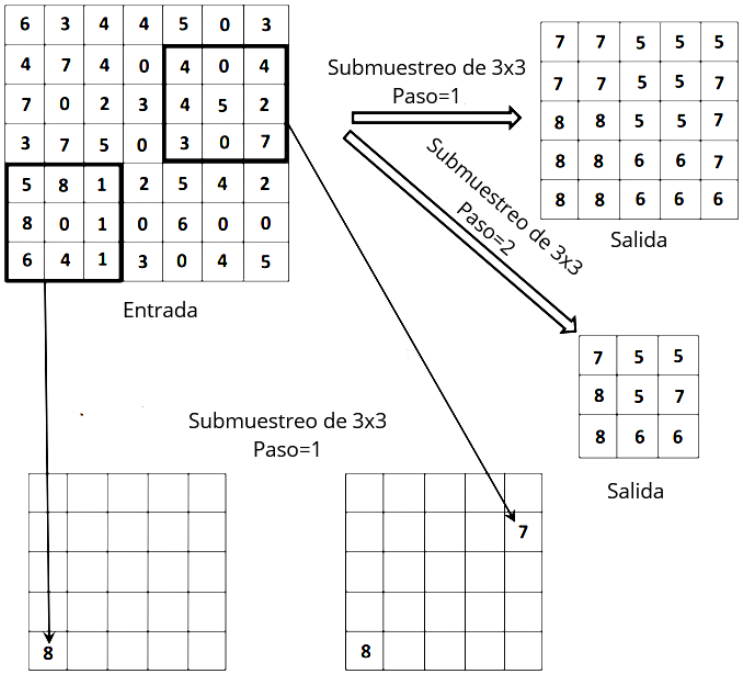


Figura 2.21: Ejemplo de un *Max-Pooling* con una región de *pooling* de 3×3 [Aggarwal C., 2018]

$$O_w = \frac{W - F_w + 2P}{S_w} + 1 \tag{2.4}$$

donde O_w es la dimensión de la salida (ya sea largo o ancho), W es la dimensión de la imagen original, F_w es el tamaño del filtro, P es el tamaño de padding y S_w es el tamaño de paso.

Capa de submuestreo: Otra capa importante que se usa en las CNNs es la capa de submuestreo (*pooling*). El submuestreo o *pooling* disminuye la dimensión espacial de la entrada a través de una reducción de parámetros. Esta capa no tiene ningún peso sináptico. Ademas, los parámetros que se ocupan son el tamaño del filtro ($n \times n$) y el paso, ya que las operaciones se realizan con la información que se tiene disponible en la entrada. Esta capa tiene el efecto de una segunda extracción de características, cuyo objetivo principal es reducir las dimensiones de los mapas de características y aumentar la robustez de la extracción. Es una práctica común agregar una capa de submuestreo entre dos capas convolucionales. Existen dos operaciones típicas de submuestreo: el submuestreo máximo y el submuestro promedio. La primera operación toma el valor máximo encontrado en el filtro (ver Figura 2.21), mientras que la segunda calcula el promedio de los valores en un filtro.

Es importante considerar el tamaño del filtro, ya que el submuestreo con grandes campos receptivos (es decir, un tamaño de filtro grande) es muy destructivo debido a la cantidad de información que se elimina. Con la ecuación 2.5 se puede calcular el tamaño del mapa de características de salida a partir del tamaño del mapa de características de entrada.

$$O_w = \frac{I_w - F_w}{S_w} + 1 \tag{2.5}$$

donde O_w es la dimensión de la salida (ya sea largo o ancho), I_w es la dimensión de la entrada, F_w es el tamaño del filtro, S_w es el tamaño de paso.

Capa completamente conectada: Las capas completamente conectadas forman un perceptrón multicapa. En la Figura 2.22, se aprecia la arquitectura de una CNN estándar, en la cual se observa cómo las dimensiones a lo largo y ancho de la imagen de entrada se van reduciendo, mientras que la profundidad (es decir, los mapas de características) se va incrementando. De igual forma, se aprecia que la capa completamente conectada tiene como entrada a cada elemento de la última capa antes de ésta. Ésto se consigue a través de un aplanado, el cual consiste en transformar un arreglo tridimensional a uno de una dimensión, concatenando toda la información del arreglo de tres dimensiones.

Capa de salida: Es la decisión final de clasificación de la red. Normalmente es representada como un vector de probabilidades cuya suma es igual a la unidad, lo anterior representa la función de activación vista anteriormente softmax.

Las partes antes mencionadas son las principales que componen una CNN de manera general, sin embargo, dentro de las CNN existen diversas arquitecturas, en la siguientes subsecciones se profundizará en algunas de estas, las cuales se utilizan

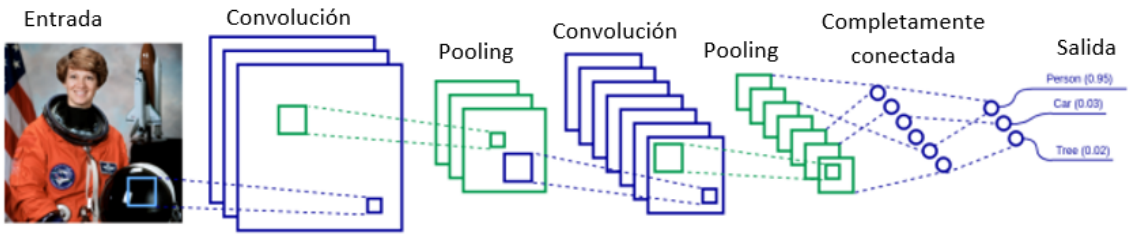


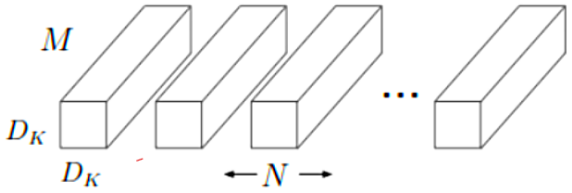
Figura 2.22: Estructura de una CNN estándar, las capas convolucionales y completamente conectadas se muestran en color azul, mientras que las de *pooling* en verde [Vasilev et al., 2019]

en este proyecto de tesis.

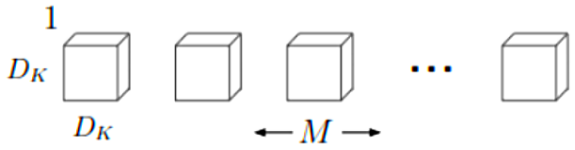
MobileNet

Esta arquitectura fue elaborada para el concurso ImageNet. Originalmente se diseñó para aplicaciones móviles y sistemas embebidos [Howard et al., 2017]. Dicha arquitectura utiliza convoluciones separables en profundidad (*depthwise convolutions*) con el fin de obtener redes neuronales más ligeras (es decir, reducir el número de pesos sinápticos a estimar). Podría decirse que estas convoluciones separables en profundidad son una forma de convoluciones factorizadas que reducen drásticamente el número de parámetros ocupados en cada capa convolucional. En la Figura 2.23 se ilustra una comparación entre una CNN normal y una CNN separable en profundidad. El proceso de reemplazo se describe a continuación:

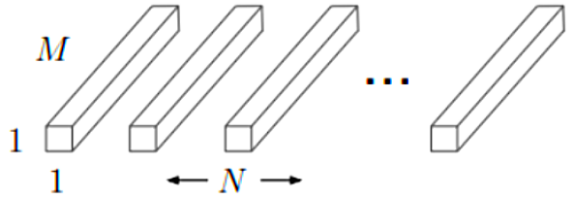
- Partiendo de varios filtros de convolución tradicional (Figura 2.23a), se tiene que M es la profundidad de entrada, D_k es el tamaño del filtro (solo se trabaja



(a) Filtro convolucional estándar



(b) Filtro convolucional en profundidad



(c) Filtro convolucional puntuall

Figura 2.23: Un filtro convolucional estándar (a) es reemplazado por dos capas: una capa convolucional en profundidad (b) y una capa convolucional puntual (c) [Howard et al., 2017]

con filtros cuadrados y de dimensión impar) y N es el número de filtros que se desean. Esta operación conlleva el cálculo de $M \times D_k \times D_k \times N$ pesos.

- En los inscisos 2.23b y 2.23c de la Figura 2.23, se puede ver que una operación convolucional tradicional se puede dividir en dos partes. En la Figura 2.23b la convolución de profundidad ocupa M filtros de dimensión $D_k \times D_k$, mientras que en el 2.23c se ocupan N filtros de una profundidad M de 1×1.

- Las Figuras 2.23b y 2.23c son una equivalencia a la Figura 2.23a, ya que al realizar las operaciones, se llega al mismo resultado de salida, solo que realizando la convolución de profundidad y puntual se ocupan $D_k \times D_k \times M + M \times 1 \times 1 \times N$ parámetros. El número anterior es una cantidad mucho menor de parámetros comparado con los ocupados por filtros de convolución tradicional.

- En el recuadro azul de la Figura 2.24 se aprecia de manera ilustrativa la equivalencia de una operación convolucional tradicional respecto a una convolución factorizada. La notación es la siguiente:

  - C representa la operación de convolución tradicional.
  - DW es la convolución de profundidad.
  - PW es la convolución puntual.
  - F es una capa densa, completamente conectada.

Un resumen detallado de las capas que forman la arquitectura MobileNet se

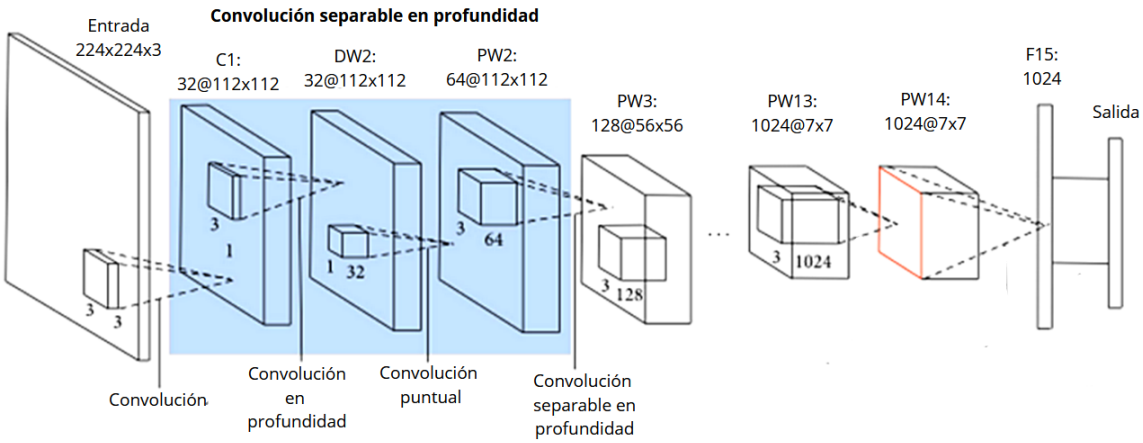


Figura 2.24: Arquitectura MobileNet [Chowdary et al., 2021]

Cuadro 2.1: Resumen de capas de la arquitectura MobileNet [Howard et al., 2017]

Tipo / Paso	**Dimensión de filtro**	**Tamaño de entrada**
Conv / s2	3 x 3 x 3 x 32	224 x 224 x 3
Conv dw /s1	3 x 3 x 32 dw	112 x 112 x 32
Conv / s1	1 x 1 x 32 x 64	112 x 112 x 32
Conv dw /s2	3 x 3 x 64 dw	112 x 112 x 64
Conv / s1	1 x 1 x 64 x 128	56 x 56 x 64
Conv dw /s1	3 x 3 x 128 dw	56 x 56 x 128
Conv / s1	1 x 1 x 128 x 128	56 x 56 x 128
Conv dw /s2	3 x 3 x 128 dw	56 x 56 x 128
Conv / s1	1 x 1 x 128 x 256	28 x 28 x 128
Conv dw /s1	3 x 3 x 256 dw	28 x 28 x 256
Conv / s1	1 x 1 x 256 x 256	28 x 28 x 256
Conv dw /s2	3 x 3 x 256 dw	28 x 28 x 256
Conv / s1	1 x 1 x 256 x 512	14 x 14 x 256
5 x Conv dw / s1 y Conv / s1	3 x 3 x 512 dw 1 x 1 x 512 x 512	14 x 14 x 512 14 x 14 x 512
Conv dw /s2	3 x 3 x 512 dw	14 x 14 x 512
Conv / s1	1 x 1 x 512 x 1024	7 x 7 x 512
Conv dw /s2	3 x 3 x 1024 dw	7 x 7 x 1024
Conv / s1	1 x 1 x 1024 x 1024	7 x 7 x 1024
Avg Pool / s1	Submuestreo 7 x 7	7 x 7 x 1024
FC / s1	1024 x 1000	1 x 1 x 1024
Softmax / s1	Clasificador	1 x 1 x 1000

muestra en el Cuadro 2.1, donde se observa que tiene como entrada una imagen de $224 \times 224 \times 3$ píxeles. Cabe mencionar que esta arquitectura se encuentra disponible en la biblioteca de acceso abierto `Keras`.

FSER

Esta arquitectura fue diseñada para realizar reconocimiento de emociones a partir de la voz [Dossou and Gbenou, 2021]. La arquitectura FSER se muestra en la Figura 2.25 y se describe de manera más específica a continuación:

- De manera general, en las capas convolucionales se utiliza la técnica de relleno cero para conservar las dimensiones de la entrada. De igual forma se utilizaron cuatro capas de submuestreo máximo de 2×2 y paso dos, con *Dropout* de 0.2 después de cada capa convolucional.

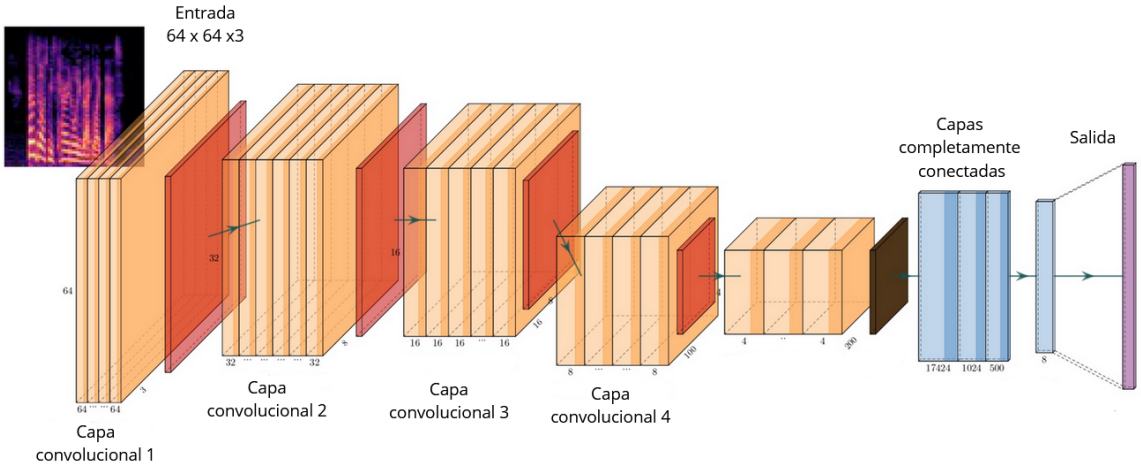


Figura 2.25: Arquitectura FSER [Dossou and Gbenou, 2021]

- La entrada es una imagen de 64×64 píxeles. En este caso se utiliza como entrada las imágenes de los espectrogramas correspondientes a las muestras de audio.

- Primera capa convolucional de ocho filtros de 5×5 y paso uno. Se utiliza la técnica de relleno con el fin de que los mapas de características generados conserven la dimensión de 64×64.

- Primera capa de submuestreo máximo. La dimensión de los mapas de características de salida es de 32×32.

- Segunda capa convolucional de 16 filtros de 5×5 y paso uno utilizando relleno. La dimensión de los mapas de características se conserva (32×32).

- Segunda capa de submuestreo máximo. La dimensión de los mapas de características de salida se reduce 16×16 píxeles.

- Tercera capa convolucional de 100 filtros de 5×5 y paso uno, con dimensión de salida de 16×16.

- Tercera capa de submuestreo máximo con dimensión de salida igual a 8×8.

- Cuarta capa convolucional de 200 filtros de 5×5 y paso uno. Se conserva la dimensión de los mapas de características de 8×8.

- Última capa de submuestreo máximo que reduce los mapas de características a una dimensión de 4×4.

- Finalmente, se realiza un aplanado de los mapas de características y conexión a una red neuronal completamente conectada, la cual consta de las siguientes subcapas:

    - Una capa densa oculta de 17,424 neuronas con función de activación ReLU.
    - Una capa densa oculta de 1,024 neuronas con función de activación ReLU.
    - Una capa densa oculta de 500 neuronas con función de activación ReLU.
    - Capa de salida con ocho neuronas y función de activación *softmax*.

Arquitecturas 1D

Las CNNs de una dimensión (1D) son un tipo de red neuronal utilizada para procesar datos secuenciales como señales de audio, datos de texto o series temporales. Estas arquitecturas funcionan de manera muy similar a las arquitecturas CNNs tradicionales; la diferencia radica en la entrada, ya que se utilizan arreglos unidimensionales. Al igual que las CNN tradicionales, estas arquitecturas utilizan filtros de convolución 1D para escanear la entrada y extraer características. La operación de convolución es típicamente seguida por capas de submuestreo, que reducen la dimensión espacial de la entrada mientras mantienen la información más importante. Finalmente, la salida de las capas anteriores se procesa por capas completamente conectadas que realizan predicciones basadas en las características procesadas [Issa et al., 2020, Zhao et al., 2019].

Autores como [Issa et al., 2020, Middya et al., 2022] han utilizado este tipo de arquitecturas de manera exitosa extrayendo las siguientes características:

- MFCC
- Espectrogramas de Mel
- Contraste espectral
- Tonnetz
- Cromagramas

Las arquitecturas 1D utilizadas en este trabajo de tesis se ilustran en la Figura

2.26 y se describen a continuación:

- **Arquitectura 1**
  - La entrada es un vector de 181 elementos conformado por diferentes características de audio: 40 MFCC, espectrogramas de Mel, tonnetz y contraste espectral.
  - Primera capa convolucional de 64 filtros de dimension 10 y paso uno. Se utiliza la función de activación ReLu.
  - Segunda capa convolucional de 64 filtros de dimension 10 y paso uno. Se utiliza la función de activación ReLu.
  - Primera capa de submuestreo máximo con dimensión de kernel igual a 8.
  - Tercera capa convolucional de 128 filtros de dimension 10 y paso uno. Se utiliza la función de activación ReLu.
  - Segunda capa de submuestreo máximo con dimensión de kernel igual a 8.
  - Capa de dropout con un factor de 0.4 y aplanado.
- **Arquitectura 2**
  - La entrada es un vector de 181 elementos formado como es descrito en la arquitectura 1.
  - Primera capa convolucional de 64 filtros de dimension 10 y paso uno. Se utiliza la función de activación ReLu.
  - Primera capa de submuestreo máximo con dimensión de kernel igual a 8.
  - Segunda capa convolucional de 128 filtros de dimension 10 y paso uno. Se utiliza la función de activación ReLu.
  - Segunda capa de submuestreo máximo con dimensión de kernel igual a 8.
  - Capa de dropout con un factor de 0.4 y aplanado.

Ambas arquitecturas se conectan a una capa densa de 128 elementos y a la capa de salida con funciónes de activación ReLu y softmax respectivamente.

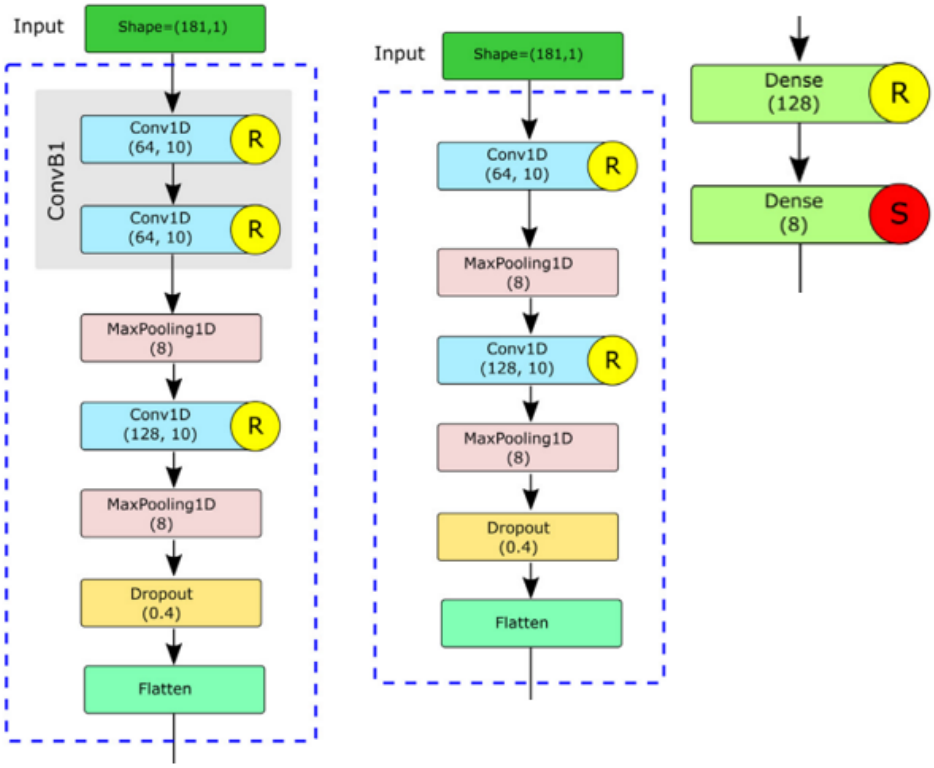


Figura 2.26: Arquitecturas 1D [Middya et al., 2022]

2.5.3. Mejoras en el desempeño de las CNNs

Con el fin de mejorar el desempeño de clasificación de las CNNs, han surgido diversas técnicas, dentro de las cuales se pueden mencionar las siguientes [Bhardwaj et al., 2018, De Marchi and Mitchell, 2019]:

- Preprocesamiento de información. El preprocesamiento normalmente se utiliza con el fin de limpiar la información, eliminar el ruido y realzar características. Esta técnica es externa a las CNNs, ya que no depende de ningún parámetro o hiperparámetro de red.

- Decaimiento de pesos. El decaimiento de pesos consiste en reducir todos los pesos a valores pequeños, debido a que pesos grandes conllevan a errores grandes, por lo cual esta técnica busca reducir el error de clasificación.

- *Dropout*: Esta técnica desactiva, aleatoriamente y cada cierto tiempo, neuronas en todas las capas de la red, con el fin de asegurar que estas aprendan las características más útiles y representativas de los elementos a clasificar.

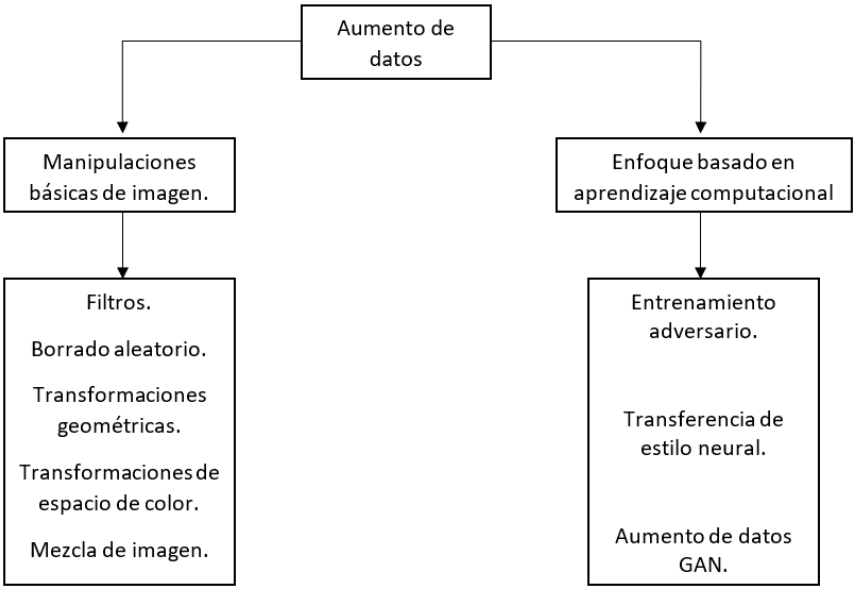


Figura 2.27: Categorización de técnicas para aumento de datos [Wang et al., 2017]

- Aumento de datos: Esta técnica consiste en incrementar los datos disponibles de entrenamiento para evitar el sobre-entrenamiento.

De manera más profunda, el aumento de datos es el proceso de incrementar la cantidad y diversidad de los datos. Esta técnica se ocupa cuando el conjunto de datos de entrenamiento es muy pequeño o para evitar sobre-entrenamiento. A diferencia de otras técnicas como *dropout*, pre-entrenamiento o normalización por lote, el aumento de datos se basa en la raíz del problema: el conjunto de datos de entrenamiento [Shorten and Khoshgoftaar, 2019].

Una de las ventajas del aumento de datos es que no se necesitan recopilar miles o millones de datos de manera práctica, ya que aumenta el tamaño del conjunto de datos e introduce variabilidad a éste transformando los datos que ya se tienen. El aumento de datos se puede realizar mediante dos maneras [Wang et al., 2017, Shorten and Khoshgoftaar, 2019]:

- La deformación de datos: En esta técnica, la imágenes son transformadas manteniendo la etiqueta de clase, ya sea mediante transformaciones geométricas, de color, borrado aleatorio, etc.

- Sobremuestreo de datos: Técnica en la que se crean instancias sintéticas y se añaden al conjunto de datos. Algunas técnicas son mezcla de imágenes, aumento de espacio de características y uso de GANs.

En la Figura 2.27 se aprecia una categorización de las técnicas de aumento de datos existentes [Wang et al., 2017]. Por un lado, se tienen los métodos basados en aprendizaje profundo. Este tipo de métodos ocupa el sobremuestreo de datos y uno de sus métodos más populares es la creación de nuevas imágenes usando GANs. Por otro lado, se encuentran las manipulaciones básicas de imagen, las cuales utilizan métodos tradicionales de aumento de datos. Este tipo de métodos ocupa la deformación de datos, donde los más populares son:

- Reflejo (*Flipping*): Gira la imagen a lo largo de la dirección horizontal o vertical.
- Acercamiento/alejamiento (*Zooming*): Consiste en acercar o alejar la imagen de acuerdo con una determinada proporción.
- Desplazamiento (*Shifting*): Cambia o desplaza la imagen de cierta forma dentro del plano de la misma, por ejemplo transformando el color o realizando cambios de posición a lo largo de la imagen.
- Inyección de ruido: Consiste en perturbar aleatoriamente cada píxel de la imagen, generalmente mediante una distribución gaussiana.
- Rotación: Consiste en rotar la imagen en determinado ángulo.
- Recorte (*Cropping*): Nos permite recortar o seleccionar un área en particular de una imagen.
- Cambios de brillo y contraste.

Capítulo 3

Desarrollo del proyecto

En este capítulo, se describe el hardware y software implementados en los módulos del proyecto para el reconocimiento de emociones. De igual manera, se presentan las metodologías empleadas, tanto para la construcción de la base de datos de emociones en español mexicano de la Universidad Tecnológica de la Mixteca, denominada UTeMo, como para el modelo a implementar. Además, se presenta una descripción de los módulos que conforman el método de clasificación de emociones híbrido usado en este trabajo.

3.1. Especificaciones de hardware y software

Los experimentos realizados con la finalidad de implementar un nuevo método híbrido para el reconocimiento de emociones descritos en este proyecto de tesis, se implementaron en un servidor con procesador Intel(R) Xeon(R) CPU ES-2620v4 a 2.1GHz x 32, 32 GB de RAM y una GPU Nvidia Quadro M4000. La plataforma utilizada fue el sistema operativo Ubuntu 22.04 LTS de 64 bits.

Para el desarrollo de la presente tesis, el lenguaje de programación utilizado fue `Python 3.9` y las bibliotecas de acceso abierto usadas en conjunto con este lenguaje son: `Keras`, `Tensorflow`, `Librosa`, `Moviepy`, `OpenCV` y sus respectivas dependencias, así como el programa de libre acceso `PRAAT`. `Python` permite crear bibliotecas y ofrece herramientas multi-propósito y `PRAAT` permite la implementación de software para análisis acústico. De igual forma, para el cómputo paralelo se emplea `CUDA 11.5` con los drivers de Nvidia correspondientes.

3.2. Construcción de la base de datos

En el contexto de reconocimiento de emociones multimodal, a partir del análisis de voz y expresiones faciales, existen tres tipos de bases de datos [Nerio et al., 2018, Wu et al., 2014, Avots et al., 2019]: las bases de datos actuadas [Livingstone and Russo, 2018], las bases de datos inducidas o semi actuadas [Nerio et al., 2018] y las bases de datos naturales o espontáneas [Perez Rosas et al., 2013].

En las bases de datos actuadas, las interpretaciones son realizadas por actores o personas capaces de interpretar directamente una emoción predefinida; en las bases inducidas o semi-actuadas, se busca inducir la emoción al intérprete mediante recursos audiovisuales; en las bases de datos naturales, se captura la expresión de la emoción en el momento exacto en la que se presenta, por esta razón este tipo de base de datos es más difícil de realizar. Debido a que las bases de datos actuadas se prestan para obtener grabaciones de calidad ya que se realizan en un entorno controlado, en esta tesis se opta por construir una base de datos de este tipo.

La construcción de la base de datos es necesaria debido a que no existe una base de datos pública que se adecúe al lugar y al contexto en el que se realiza la investigación (México). La mayoría de bases de datos para el reconocimiento de emociones se han construido para el idioma inglés [He et al., 2020, Rahdari et al., 2019] y otras pocas se han grabado en Castellano [Barra-Chicote et al., 2008, López et al., 2006]. Sin embargo, no hay un conjunto de datos de emociones disponible en español de México que contenga las seis emociones básicas, las cuales son: tristeza, sorpresa, felicidad, ira, miedo y asco, además de un estado neutral [Ekman et al., 1983]. El requisito del idioma (español mexicano) es necesario debido a que la expresión de emociones mediante el habla carece de universalidad, a diferencia de las expresiones faciales que sí son universales [Darwin, 1872, Mora Teruel and Sanguinetti, 2004].

Tomando en cuenta el idioma y las seis emociones antes mencionadas, además de un estado neutral, se propone una base de datos llamada UTeMo, haciendo alusión al juego de palabras entre *emo* de emoción y *UTM* por la Universidad Tecnológica de la Mixteca. UTeMo se construyó bajo el esquema que se muestra en la Figura 3.1, el cual consta de los siguientes pasos:

1. Obtener un marco de referencia respecto a otras bases construidas para el reconocimiento de emociones, el cual consiste en analizar las diversas configuraciones de escenario, frases y personas (actores), con el fin de implementar una configuración adecuada.

2. De acuerdo con la información recopilada en el paso anterior, se seleccionaron

Figura 3.1: Esquema para la construcción de la base de datos.

las frases en español con base en la metodología de [Nerio et al., 2018], la cual propone frases acordes al estado emocional.

3. Respecto a la configuración del escenario, se utiliza la configuración presentada en la Figura 3.2, la cual está basada en [Livingstone and Russo, 2018], debido a que utiliza poco equipo especializado sin perder calidad en las grabaciones.

4. Tomando en consideración la construcción de la base de datos actuada de [Wu et al., 2014], se ha solicitado la ayuda de varios actores para las grabaciones. Este tipo de bases de datos asegura la correcta expresión de la emoción y disminuye el tiempo de grabación.

5. Una vez realizadas las grabaciones, se realiza el etiquetado de emociones a cada una de las frases.

6. Por último, se realizan las validaciones: dos subjetivas (cualitativas) y una

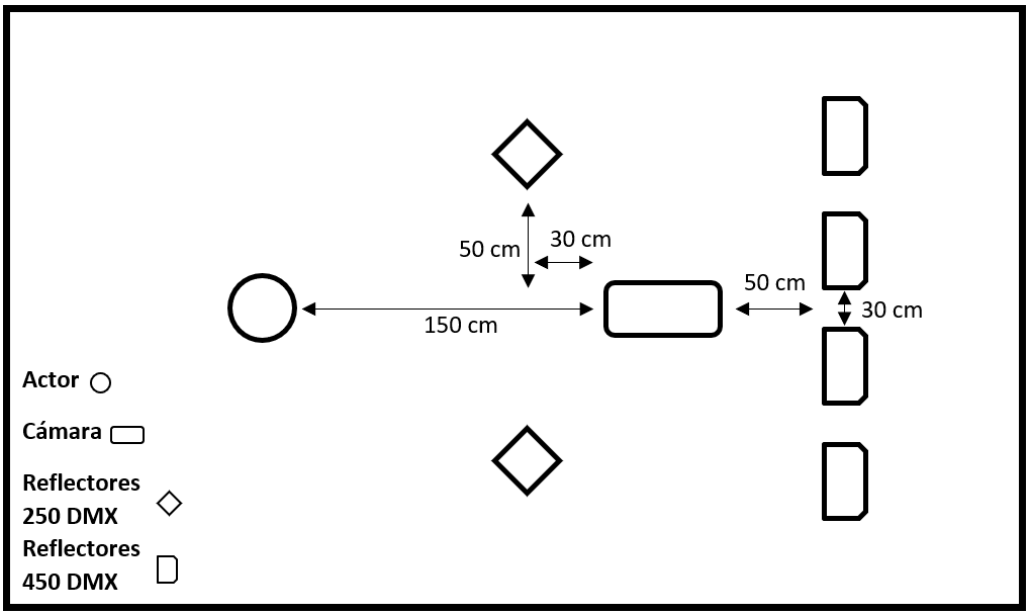


Figura 3.2: Configuración empleada para las grabaciones

objetiva (cuantitativa):

- Validación subjetiva: La primera de éstas consiste en que cada muestra sea validada por un experto en el tema de las emociones, asegurándose de que las interpretaciones de los actores se realicen correctamente y se descarten aquellas muestras mal interpretadas. La segunda consiste en aplicar una encuesta a 28 personas, a las que se les pidió que emitieran dos juicios por cada muestra: uno de tipo categórico y otro relacionado con la intensidad emocional.

- Validación cuantitativa: Esta validación consiste en utilizar métodos tradicionales de aprendizaje computacional para evaluar la base de datos. Esto sirve para corroborar la viabilidad de la base de datos propuesta para tareas de clasificación.

Para la construcción de la base de datos propuesta, participan diferentes actores pertenecientes a grupos de teatro, tanto públicos como privados, quienes radican en la Heroica Ciudad de Huajuapan de León, Oaxaca. Se cuenta con un total de 14 personas, 7 mujeres y 7 hombres, con un promedio de edad de 27 años.

Las emociones que son tomadas en consideración para la base de datos son: ira, sorpresa, felicidad, miedo, asco y tristeza, además de un estado neutral. Estas emociones pertenecen al modelo usado por Paul Eckman [Ekman, 1992] que fue descrito

en la Sección 2.1. Las frases utilizadas para la construcción de la base de datos se encuentran en los Cuadros 3.1 y 3.2. El número de muestras varía entre emociones, debido a que son seleccionadas por diversos autores para realizar reconocimiento de emociones. En [Nerio et al., 2018], no se considera la emoción o estado neutral (Cuadro 3.2), por lo cual las frases para este estado se toman de otras bases de datos (Cuadro 3.1) [Livingstone and Russo, 2018, Haq and Jackson, 2010, Burkhardt et al., 2005].

Para garantizar la correcta recolección de los datos, se adopta el modelo de la base de datos RAVDESS [Livingstone and Russo, 2018] que posee una distribución del equipo de grabación semejante a la Figura 3.2, donde se sugiere que las interpretaciones sean realizadas por actores. Cabe mencionar que, cada muestra de la base de datos consiste en un video de un actor interpretando una frase de los Cuadros 3.2 y 3.1 en su correspondiente emoción.

El espacio de trabajo, así como el equipo utilizado, son facilitados por la Universidad Tecnológica de la Mixteca. El entorno de trabajo consta de una habitación acústicamente aislada con fondo blanco. El equipo de grabación utilizado es una cámara de video CANON modelo XF105a, dos reflectores Fluotec 250DMX y cuatro reflectores Fluotec 450DMX. El audio es grabado con el micrófono interno de la cámara, el cual es un micrófono de condensador estéreo.

Una vez capturadas las muestras, se ordenan por emoción; posteriormente, se realiza una conversión al formato AVI (por las siglas en inglés de *Audio Video Interleave*). Las muestras etiquetadas siguen el siguiente formato para el nombre del archivo: `Emocion_Actor/Actriz`N`_P`M `_F`L, donde la N, representa el número de actor/actriz, M es el número de iteración (en un rango de 1 a 3) y L es el número de frase. Por ejemplo, una etiqueta para el nombre del archivo sería: `Asco_Actriz1_P2_F5.avi` que indica que en ese archivo se está interpretando la

Cuadro 3.1: Frases para el estado neutro o emoción neutra.

Base de datos	**Frase**
Emod-DB [Burkhardt et al., 2005]	El mantel está sobre la nevera
	En siete horas estará listo
	La hoja de papel negra se encuentra ahí arriba
	¿Qué pasa con las bolsas que están debajo de la mesa?
	Esta en el lugar donde lo guardamos siempre
RAVDESS [Livingstone and Russo, 2018]	Los niños estan sentados junto a la puerta
	Los perros estan sentados junto a la puerta
SAVEE [Haq and Jackson, 2010]	La mejor forma de aprender es resolviendo problemas extra
RML [Wang and Guan, 2008]	Se agradece una pronunciación clara
	Ella es mas delgada que yo.

Cuadro 3.2: Frases por emoción tomadas de [Nerio et al., 2018]

Emoción	Frase
Ira	1) ¿Qué te pasa?
	2) ¡Eso a mí que me importa!
	3) ¡O te vas o te boto!
	4) ¿Me vas a atender o no?
	5) ¿Sabes qué? ¡déjalo así!
	6) ¡No me molestes!
Sorpresa	1) ¡No puede ser! ¿en serio?
	2) ¡Qué! ¡yo no sabía eso!
	3) ¡Jamás lo hubiera creído!
	4) ¡No me lo esperaba!
	5) ¡No te creo! ¿de verdad?
	6) ¿De verdad? ¡No sabía!
	7) ¿Es en serio?
Felicidad	1) ¡Gané!
	2) ¡Que genial! ¡pasé!
	3) ¡No me lo creo! ¡qué suerte!
	4) !Lo logré¡ ¡al fin!
	5) ¡No puede ser! ¡qué bien¡
	6) ¡No lo creo! ¡funciona!
Miedo	1) ¡No!, ¡no me hagas daño!
	2) ¡Ya no tengo más!, ¡no tengo nada!
	3) ¡No!, ¡no me robes!
	4) ¡Aléjate!, ¡Aléjate!
	5) ¡Aléjate por favor!
	6) ¡No!, ¡por favor!
Asco	1) ¡Esto sí está feo!
	2) ¿Qué hay en el plato?
	3) !Qué repugnante¡
	4) ¿Qué asco? ¿qué es esto?
	5) ¡Un bicho!
	6) ¿Qué es esto?
Tristeza	1) Todo iba tan bien, ¡no sé qué pasó!
	2) ¡Lo/la extraño, pero se fue!
	3) ¡Ya no será lo mismo!
	4) ¡Dime que no es verdad!
	5) ¡Aún sentía algo por él/ella!
	6) ¡Él/Ella fue parte de mi vida!
	7) ¡No pude hacerlo!

emoción de asco, la persona grabada es la Actriz1, es la segunda iteración y la frase que se está actuando es la número 5.

Posterior al etiquetado de los datos, se valida cada una de las muestras con la psicóloga Ivette Jiménez García, quien es especialista en neurociencias y maestrante (al momento de la redacción de esta tesis) en psicoterapia cognitivo-conductual. Esta

validación consiste en observar cada muestra y confirmar que la expresión realizada por el actor correspondiera a la emoción que se está actuando. La segunda validación realizada consistie en una encuesta en la que participaron 28 personas; éstas se aplican a través de la plataforma Google Forms. Es necesario mencionar que, únicamente se ocupa el diez por ciento de la base de datos (un subconjunto estratificado representativo) dividido en ocho subconjuntos, proporcionando subconjuntos diferentes a cada uno de los participantes para que validen cada muestra emitiendo dos tipos de juicios:

1. Juicio categórico: Cada persona puede elegir entre 8 diferentes opciones, una por cada estado emocional además de una opción extra con la etiqueta *ninguna de las anteriores*, con el fin de validar que la emoción interpretada corresponda a su categoría.

2. Juicio de intensidad: Por cada muestra, se solicita a cada persona que eligiera un valor entre 1 y 3 donde 1 denota una intensidad de expresión *débil*, 2 *normal* y 3 *fuerte*.

Para realizar una interpretación adecuada de los resultados de las encuestas realizadas, se utilizó una medida estadística conocida como el Factor Kappa de Fleiss [Fleiss et al., 1981]. Gracias a este factor, es posible conocer el nivel de concordancia que existe entre los juicios categóricos realizados por las personas y la etiqueta original de cada muestra realizada por el actor. Según los autores, un factor Kappa cercano a 1,0 es ideal; en la práctica, dado que este análisis se basa en encuestas, un factor Kappa mayor o igual a 0,75 es excelente.

La última validación que se realiza es una validación objetiva mediante algoritmos tradicionales de aprendizaje automático en los que es necesario un preprocesamiento adicional para extraer características útiles del audio y video.

Esta validación comienza con un preprocesamiento aplicado a la voz y a las imágenes de las expresiones faciales. Primeramente, el preprocesamiento de audio consiste en lo siguiente:

1. Extraer el audio de cada muestra de video.

2. Amplificar el audio en 10dB.

3. Eliminación de silencios inicial y final de cada muestra.

Posteriormente, el preprocesamiento de imágenes consiste en:

1. Extraer una imagen para cada muestra de video.

2. Aplicar una ecualización CLAHE (ver Sección 2.2.1).

3. Extraer un rostro de la imagen con el algoritmo Viola-Jones [Viola and Jones, 2004].

Cabe mencionar que la imagen extraída se selecciona de un fotograma de video correspondiente a $T = 2s$, ya que éste es el tiempo medio de la mayoría de las muestras. Para muestras más cortas, se selecciona la imagen del fotograma correspondiente a $T = 1s$.

La extracción de las características de voz se realiza a partir de las muestras de audio preprocesadas. En este caso, la herramienta que se utiliza es el *software* libre `librosa`, usada comúnmente para el análisis de música y audio. Con esta herramienta, se pueden calcular características prosódicas y espectrales, las cuales son los dos tipos más utilizados en la literatura de clasificadores convencionales para reconocimiento de emociones. Dentro del grupo de características prosódicas se calcula la Tasa de Cruce por Cero (ZCR, por sus siglas en inglés), el Centroide y la Energía RMS, extrayendo una sola característica por muestra. Para las características espectrales, se realiza el cálculo de los MFCC (ver Sección 2.3.2), los cuales están representados por 13 vectores de diferente dimensión dependiendo de cada muestra. Con el fin de obtener un número de características constante, se procede a calcular la media y desviación estándar de estos vectores, obteniendo así 26 características por cada muestra. Además, se ha calculado la primera y segunda derivada de los MFCC, obteniendo 52 características espectrales adicionales. Finalmente, se obtienen otras 17 características espectrales a partir de los Coeficientes de Predicción Lineal (LPC, por sus siglas en inglés). En total, se obtienen 98 características prosódicas y espectrales, cuya distribución se muestra en el Cuadro 3.3.

Las características de las imágenes que corresponden a las expresiones faciales se

Cuadro 3.3: Distribución por muestra de las características de audio.

Tipo	**Característica**	**Número de características**
Prosódicas	Radio de cruce por cero	1
	Centroide	1
	Energía	1
Espectrales	Coeficientes Cepstrales de Mel (MFCC)	26
	Primer derivada MFCC	26
	Segunda derivada MFCC	26
	Coeficientes de Predicción Lineal	17
	Total	98

obtienen con ayuda de las bibliotecas de *software libre* `SciKit-Learn`, `SciKit-Image` y `OpenCV`. Para este primer experimento se utiliza una imagen extraída por muestra en escala de grises. Estas imágenes tienen una dimensión de 256×256 píxeles, las cuales son aplanadas en un solo vector de 65536 elementos por imagen. Aunque en una sola imagen no se puede detectar movimiento de una expresión facial, y por consecuencia la emoción correspondiente, se realiza el experimento para encontrar las características sencillas que logran capturar la información (limitada) de las emociones, ya que para esta fase interesa validar la base de datos construida para tareas de clasificación. El primer conjunto de características se forma a partir de los Histogramas de Gradientes Orientados (HOG, por sus siglas en inglés), usando un tamaño de ventana de 24×24 píxeles obteniendo así un total de 800 características por muestra. De igual forma, el segundo conjunto de características se forma a partir de los Patrones Binarios Locales (LBP, por sus siglas en inglés), usando una ventana de 24 × 24 píxeles, obteniendo un total de 65536 características por muestra, las cuales son transformadas con PCA hasta para obtener 433 características. Finalmente, el tercer conjunto de características consta de 900 características de la variante *Uniform* de los LBPs. A manera de resumen, el Cuadro 3.4 muestra de forma clara la distribución de las características extraídas con base en las imágenes que corresponden a las expresiones faciales.

Para la clasificación de emociones, se realizan los experimentos de clasificación normalizando las características de manera estándar (media cero y desviación estándar unitaria). Para ello, se utilizan los clasificadores $k-$Vecinos mas cercanos (KNN), SVM y Bayes Ingenuo (NB). Para los clasificadores KNN y SVM, los parámetros adecuados en cada experimento se encuentran mediante una búsqueda de malla. En el caso de KNN se hace una búsqueda con $k = 1, 2, ..., 15$; mientras que para SVM se prueban tres kernels diferentes: lineal, polinomial y Gaussiano. Consecuentemente, la búsqueda de malla se realiza variando los parámetors $\gamma = 0{,}001, 0{,}002..,0{,}009, 0{,}015, 0{,}02, 0{,}5, 0{,}9$; y $C = 1, 2, 3, ..., 5$.

Cuadro 3.4: Distribución por muestra de las características de imagen.

Característica	**Núm de características**
Histograma de Gradientes Orientados (HOG)	800
Patrones Binarios Locales (LBP)	433
Variante Uniform de los LBP	900
Total	2133

3.3. Módulos del proyecto

Esta sección tiene la finalidad de presentar el esquema general de la presente propuesta de tesis, así como una descripción de sus módulos principales basados en métodos novedosos de aprendizaje computacional, tal como es el aprendizaje profundo. La propuesta de tesis se conforma de dos partes principales: el reconocimiento de emociones a partir de las expresiones faciales y el reconocimiento de emociones a partir de la voz. A la vez, estos módulos se combinan para formar un método híbrido para el reconocimiento de emociones.

3.3.1. Módulo de clasificación de emociones mediante las expresiones faciales

A partir de la metodología de [Zhao et al., 2018], se prueban dos enfoques para la clasificación de emociones con expresiones faciales (ver Figura 3.3). En el primer enfoque no se considera el cálculo de flujo óptico (bloque rojo de la Figura 3.3), mientras que en el segundo, sí. Lo anterior con el fin de comparar el desempeño de clasificación entre ambos enfoques y validar la vialidad del cálculo de flujo óptico para la tarea de clasificación de este proyecto de tesis.

Extracción de imágenes

El primer paso de esta metodología consiste en extraer las imágenes de las expresiones faciales directamente de las muestras de video de las bases de datos. Para ello, se ocupa la biblioteca de acceso libre `OpenCV`, debido a que esta herramienta incluye funciones especializadas para la extracción de imágenes a partir de un video.

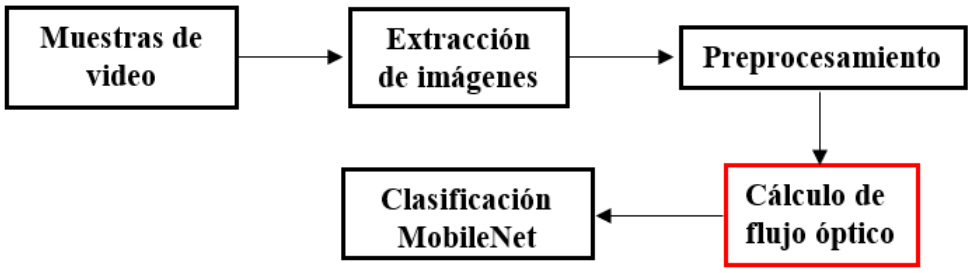


Figura 3.3: Metodología para el reconocimiento de emociones mediante las expresiones faciales.

De esta manera, se proponen dos extracciónes de imágenes: la primera extracción de imágenes se utiliza para realizar el cálculo de flujo óptico. Este cálculo consiste en extraer 8 cuadros (imágenes) de cada video distribuidos de manera uniforme en el tiempo, las imágenes primera y última representan el primer y último cuadro del video en cuestión. En esta extracción, se toman en cuenta estas 8 imágenes debido a que el cálculo de flujo óptico permite ver los micro y los macro movimientos entre dos imágenes. En la segunda extracción, solo se toman en cuenta 5 de estas 8 imágenes (ver Figura 3.4), descartando las primeras dos y la última de la extracción anterior por dos razones: la primera es que en las imágenes consideradas se aprecia una expresión facial más representativa, la segunda razón es que las imágenes descartadas son muy similares al estado neutral, por lo que pueden causar confusión en la clasificación.

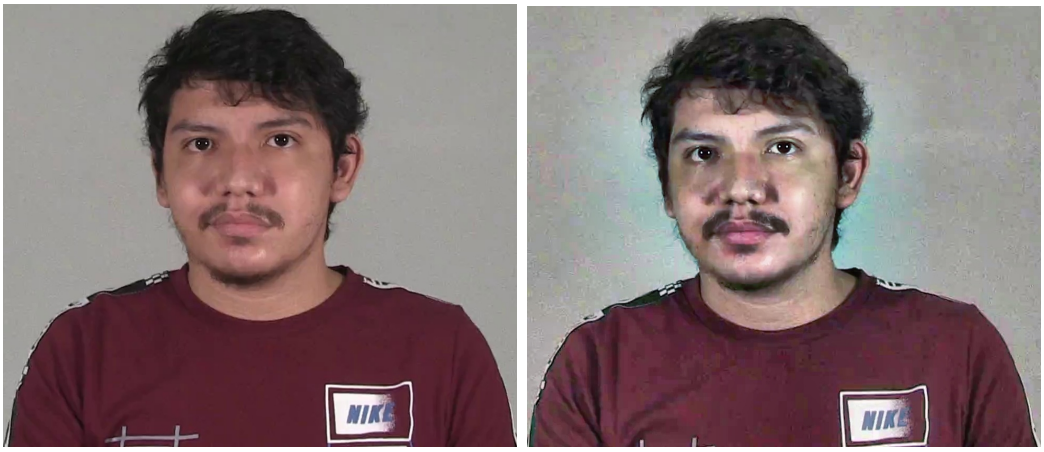
Preprocesamiento

Como se menciona en la sección 2.2.1, el preprocesamiento en imágenes tiene dos objetivos [Gonzales C. and Woods E., 2008]. El primero es eliminar información no deseada, como el ruido; y el segundo es resaltar información útil para la clasificación. Para esta tarea, se seleccionan dos métodos de preprocesamiento: el método de Ecualización de Histograma Adaptativa Limitada por Contraste (CLAHE, por sus siglas en inglés) y el algoritmo Viola-Jones.

En este caso, se selecciona CLAHE debido a que se desea destacar los rasgos faciales en todas las muestras y este método ha demostrado tener buen desempeño de acuerdo con la literatura [Hossain and Muhammad, 2019]. Este método hace que los valores de luminancia de cada imagen se distribuyan en un rango de 0 a 255 (o muy cercano a este rango), lo cual significa que la gama de luminancia se encuentra repartida de mejor forma en las imágenes ecualizadas con CLAHE y, por lo tanto, haya un mayor realce de características. En la Figura 3.5 se puede apreciar una comparativa entre dos imágenes que ilustran esta técnica. Cabe mencionar que este método se implementa con ayuda de la biblioteca de software libre `OpenCV`, importada desde el lenguaje de programación `Python`.

Con el fin de eliminar el fondo y descartar información no relevante en la imagen,

Figura 3.4: Selección de imágenes por muestra de video.



(a) Imagen original (b) Imagen ecualizada con CLAHE

Figura 3.5: Comparación de la imagen original y el preprocesamiento utilizado.

es decir, para extraer la región de interés, se aplica el *algoritmo Viola-Jones* [Viola and Jones, 2004]. Para fines de este trabajo de tesis, se utiliza la versión preentrenada de este algoritmo también disponible en `OpenCV`.

Para la implementación del algoritmo Viola-Jones se siguen los pasos siguientes:

1. Se realiza la lectura de la imagen y se crea una copia de ésta en escala de grises, ya que el algoritmo Viola-Jones original sólo funciona con este tipo de imágenes.

2. Sobre la copia realizada, se le aplica el algoritmo Viola-Jones y se obtienen las coordenadas de la ubicación del rostro en la imagen.

3. Con las coordenadas obtenidas, se procede a extraer el rostro en la imagen original a color.

En la Figura 3.6 se observa una imagen completa y el cuadrado rojo enmarcando el rostro detectado por el algoritmo Viola-Jones. Asimismo, se observa la ventana emergente que aparece en la parte derecha superior que muestra la extracción del rostro de la imagen original, es decir la *Región de Interés* (RoI).

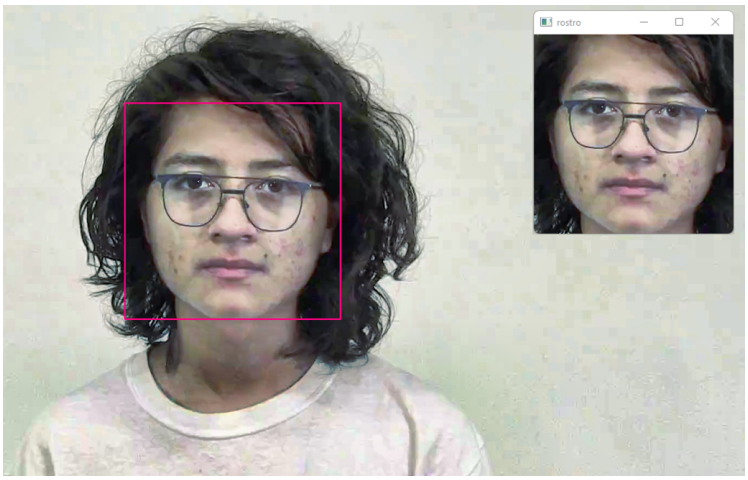
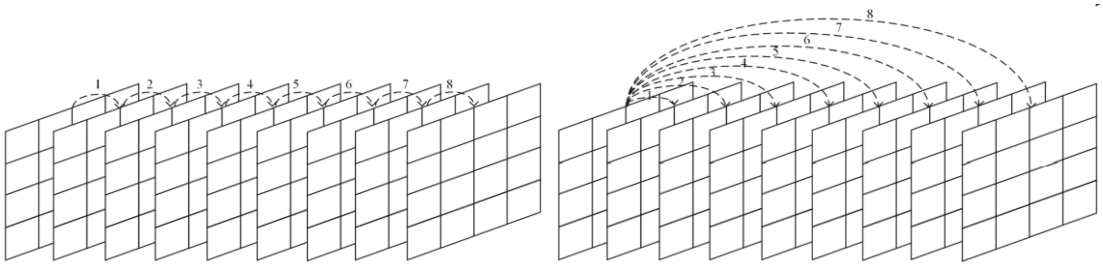


Figura 3.6: Detección de rostro con el algoritmo Viola-Jones.

Cálculo de flujo óptico

El cálculo de flujo óptico (bloque rojo de la Figura 3.3) se utiliza para análisis de micro y macro movimientos en el rostro. Con el cálculo de flujo óptico se obtienen nuevas imágenes que representan el movimiento entre imágenes consecutivas; estas imágenes se pueden utilizar en la clasificación en lugar de las imágenes de las expresiones faciales. Esta tarea se realiza con la herramienta `OpenCV` con la cual se puede calcular el flujo óptico entre dos imágenes seleccionadas de dos formas: La primera forma consiste en calcular el flujo óptico entre imágenes consecutivas, es decir, la imagen n con la imagen $n+1$ (Figura 3.7a); la segunda forma es realizar el cálculo de esta operación entre la imagen 1 y la imagen n (Figura 3.7b).



(a) (b)

Figura 3.7: Métodos de selección de pares de imagen para el cálculo de flujo óptico [Zhao et al., 2018]

Figura 3.8: Distribución de la información de las bases de datos.

Clasificación

Para la clasificación con expresiones faciales se hace una comparación entre las arquitecturas MobileNet, ResNet, VGG16 y una 3DCNN (implementadas con ayuda de la biblioteca de software libre `Keras`), de las cuales la arquitectura MobileNet propuesta en [Howard et al., 2017] y descrita en la Sección 2.5.2 ofrece un mejor desempeño. Posteriormente se realiza una búsqueda de malla gruesa con el fin de encontrar los hiperparámetros adecuados en tres bases de datos: UTeMo, SAVEE y RAVDESS. Estos experimentos se realizaron utilizando la distribución de la Fig. 3.8, con el conjunto *Train* y *Val* para encontrar los hiperparámetros y el conjunto de *Test* para probar el desempeño final del modelo.

La búsqueda de malla se realizó variando los siguientes hiperparámetros:

- Número de épocas: [10, 15, 25, 50, 100]
- Optimizadores: [Adam, SGD]
- Tamaño de lote: [32, 50, 64]
- Inicializadores de pesos sinápticos: [ImageNet, Random, HeNormal y Glorot-Normal]

3.3.2. Módulo de clasificación de emociones mediante la voz

Basándonos en los resultados presentados en [Dossou and Gbenou, 2021] y [Middya et al., 2022], se propone una metodología similar (ver Figura 3.9) que consta de cinco pasos partiendo de las muestras de audio. Ésta se explica a continuación.

Muestras de audio

Dado que la principal fuente de información de las bases de datos es audiovisual, es decir, se encuentra en formato de video, es necesario separar ambos canales de

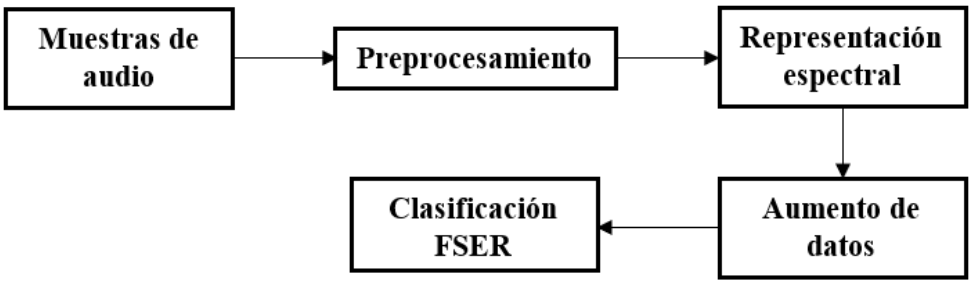


Figura 3.9: Metodología para el reconocimiento de emociones mediante la voz.

información (audio e imagen). Para esta tarea, se utiliza la herramienta de software libre `Moviepy`, disponible para el lenguaje de programación `Python`, con la cual se extrae el audio de cada muestra de video en formato WAV, debido a que este tipo de archivo mantiene una mejor calidad de audio.

Preprocesamiento

Como se menciona en la Sección 2.3.1, el preprocesamiento de audio es necesario debido a que las muestras generalmente contienen información no deseada como el ruido censado del ambiente durante la grabación [Koduru et al., 2020], el cual puede eliminarse usando filtros. Para fines de este proyecto, se parte del análisis de las grabaciones obtenidas en la base de datos UTeMo y se realiza la eliminación del ruido ambiental mediante el método de *muestreo de frecuencias*, así como el recorte de fragmentos de audio donde no está presente la voz.

La eliminación de ruido ambiental permite eliminar información no deseada sin afectar significativamente la información vocal. La eliminación de silencios beneficia el coste de procesamiento, ya que los silencios inicial y final no aportan información relevante para la clasificación. Finalmente, la amplificación de 10 dB sirve para realzar características. En la Fig. 3.10 se muestra el efecto del preprocesamiento en el espectro temporal de dos muestras de audio, donde el espectro superior corresponde al audio original y el espectro inferior al audio (muestra) resultante.

Representación espectral

Una vez que la información de audio de todas las muestras están preprocesadas, la información se transforma de audio a imagen mediante una representación espectral. Esta transformación se realiza debido a que se pueden extraer atributos de

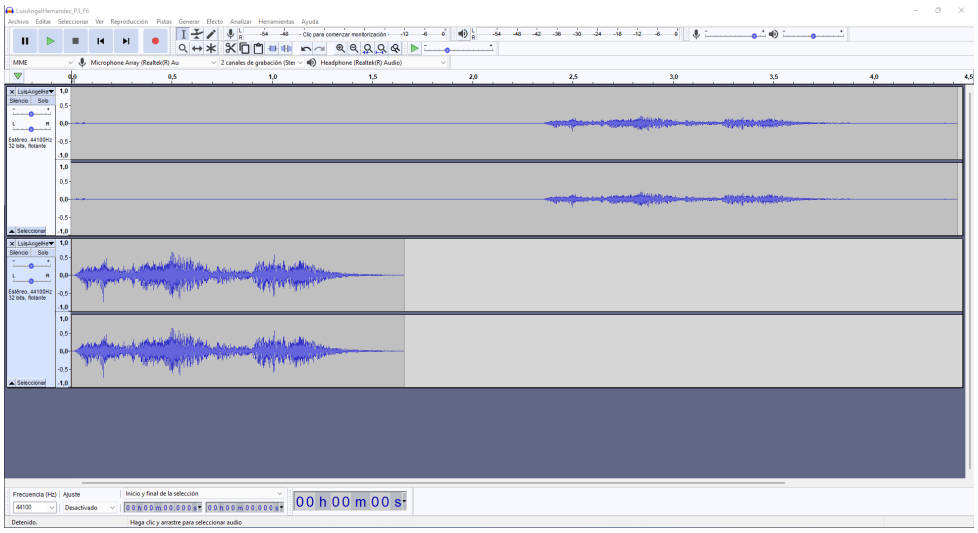


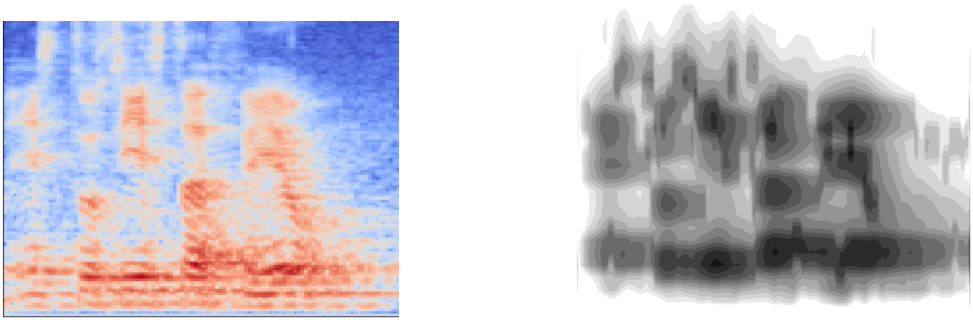
Figura 3.10: Herramienta Audacity. En la parte superior de la imagen se muestra el espectro temporal antes del preprocesamiento, mientras que en la parte inferior se muestra el equivalente de la muestra de audio preprocesada.

manera más sencilla en comparación con una representación en el dominio temporal. El análisis espectral se enfoca en la representación del espectro de audio a partir de determinadas transformaciones [Pérez Espinosa and Reyes García, 2010].

Para obtener esta representación, se utilizan dos herramienta de software libre: `Librosa`, disponible para el lenguaje de programación `Python`, y PRAAT, que es un programa multiplataforma de código abierto diseñado en la Universidad de Amsterdam por Paul Boersma y David Weenink para hacer análisis fonético [Boersma and Weenink, 2021]. Con la primer herramienta, se obtienen los espectrogramas de Mel y con la segunda, los cocleogramas. Ambas representaciones espectrales generadas a partir de las muestras de audio fueron guardadas en un formato PNG de 64×64 píxeles. En la Figura 3.11 se aprecia un ejemplo de estas representaciones. Los espectrogramas se obtienen a color, mientras que los cocleogramas se muestran en escala de grises.

De igual forma, se extraen las siguientes características espectrales con el fin de implementar arquitecturas 1D en la clasificación de audio:

- Coeficientes Cepstrales de Frecuencia Mel (MFCC): 78
- Espectrograma en escala Mel: 78
- Característica de contraste espectral: 14



(a) (b)

Figura 3.11: (a) Espectrograma y (b) cocleograma generados a partir de una misma muestra de audio

- Representación Tonnetz: 12

Finalmente, las características anteriores suman un total de 182.

Aumento de datos

De manera similar a [Dossou and Gbenou, 2021], se recomienda realizar un aumento de datos a la base de datos, aplicando operaciones de reflejo, acercamiento/alejamiento y rotación. Este aumento se realiza con la clase `ImageDataGenerator` que proporciona la biblioteca de acceso abierto `Keras`.

`ImageDataGenerator` realiza las técnicas de aumento de datos configuradas mediante sus parámetros sobre una imagen y de forma aleatoria. Así pues, existe la probabilidad de que a una imagen se le apliquen todas, alguna o ninguna de las operaciones especificadas. Es importante tener esto en cuenta debido a que, el número de imágenes generadas puede variar dependiendo de cómo se defina el ciclo en que `ImageDataGenerator` se ejecute. Para este propósito, se generaron 20 imágenes nuevas por cada muestra.

Cabe mencionar que para la experimentación la base de datos se distribuye como se muestra en la Figura 3.8. Solo en el caso de las representaciones espectrales (espectrogramas y cocleogramas) se implementa la técnica de aumento de datos al conjunto *Train*. Sin embargo, para las características espectrales utilizadas en la arquitectura 1D tomadas de [Middya et al., 2022] no se aplica esta técnica.

Clasificación

La clasificación se realiza una vez que se tiene el conjunto de datos aumentado (en el caso de las representaciones espectrales). Para ello, se implementa una arquitectura llamada FSER propuesta en [Dossou and Gbenou, 2021] y las arquitecturas 1D de [Middya et al., 2022] usando la biblioteca de software libre `Keras` y descritas en la Sección 2.5.2. Se realiza una búsqueda de malla gruesa con el fin de encontrar los hiperparámetros adecuados en cada base de datos utilizada. Estos experimentos se realizan con el conjunto *Train* y *Val*. Una vez encontrada la configuración correcta, se realiza un último entrenamiento juntando ambos conjuntos (*Train* y *Val*) y probando el desempeño de clasificación usando el conjunto *Test*.

La búsqueda de malla se realiza variando los siguientes hiperparámetros:

- Número de épocas: [200, 400]
- Tamaño de lote: [16, 32, 64]
- Optimizadores: [Adam, SGD]
- Inicializadores de pesos sinápticos: [HeNormal, Glorot Normal]

3.3.3. Módulo de clasificación de emociones de forma bimodal

Como se mencionó en la Sección 2.4, en la literatura se distinguen dos enfoques para el reconocimiento de emociones mediante información audiovisual: el primero es a través de la combinación de características y el segundo, por combinación de decisiones [Trigeorgis et al., 2017]. Con la finalidad de implementar un nuevo método para reconocimiento de emociones a través de información audiovisual, se propone la metodología de la Figura 3.12 basada en la combinación de decisiones. Esta trabaja con dos módulos principales ya descritos anteriormente en la sección 3.3.2 y 3.3.1.

Para combinar las decisiones de los módulos de clasificación con voz y expresiones faciales se sigue la regla de la Ecuación 3.1, en la cual se realiza una suma ponderada de los vectores de decisión generados por cada clasificador.

$$arg\max_{i \in C}(\alpha * S_1 + (1 - \alpha) * S_2) \tag{3.1}$$

donde S_1 y S_2 son los vectores de decisión de salida provenientes de cada módulo (expresiones faciales y voz, respectivamente) y α es una constante entre 0 y 1 que funciona como ponderador.

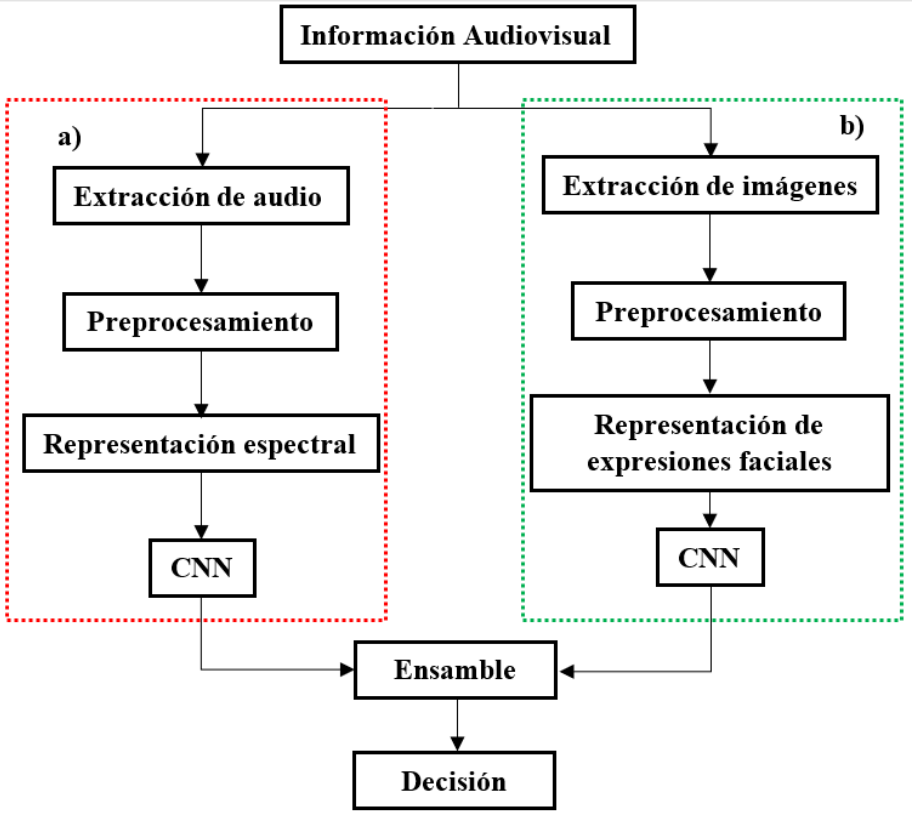


Figura 3.12: Esquema de la propuesta para reconocer emociones de forma multimodal.

En la Figura 3.13 se aprecia un ejemplo ilustrativo de la Ecuación 3.1, en donde la Fig. 3.13a representa los vectores de decisión de salida de cada clasificador, y la Fig. 3.13b ilustra la ponderación que se le asigna a cada vector. Además, la Fig. 3.13c representa la suma de ambos vectores ponderados y finalmente, la Fig. 3.13d la decisión final del clasificador.

En la implementación de este método, se pretende innovar en la parte de audio a través de la implementación de una representación espectral llamada cocleograma [Cicres, 2009]. Esta representación no ha sido utilizada con anterioridad para el reconocimiento de emociones multimodal. Por otra parte, se usaron métodos para detección de micro y macro movimientos en secuencias de imágenes, las cuales son técnicas recientemente usadas para el reconocimiento de emociones a través de información audiovisual [Deng et al., 2020].

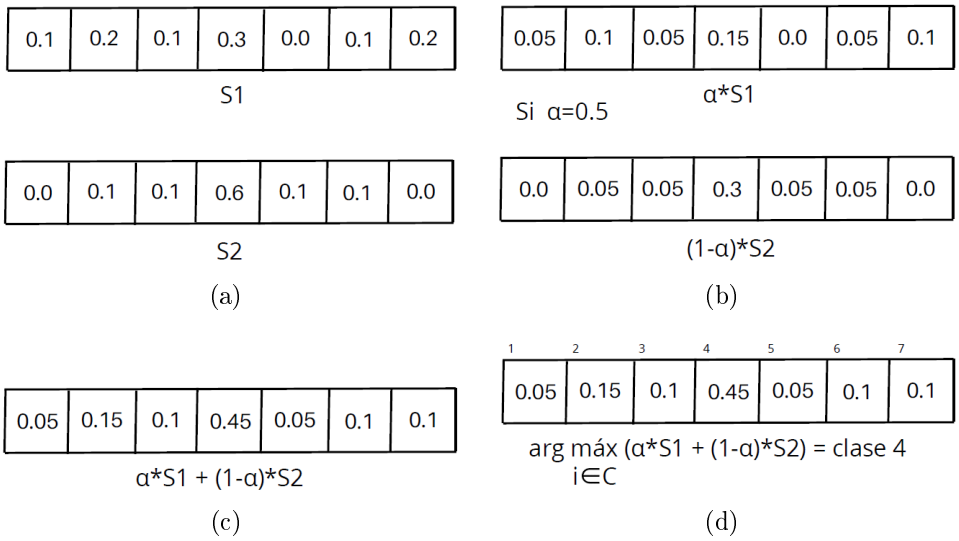


Figura 3.13: Ejemplo de uso de la Ecuación 3.1

Capítulo 4

Resultados

En este capítulo se presentan tres secciones. En la primera sección, se describen los conjuntos de datos utilizados en la experimentación, así como un análisis cualitativo de la base de datos UTeMo. En la segunda sección, se muestran los resultados obtenidos en los experimentos con métodos tradicionales de clasificación, como parte de la validación de UTeMo. Finalmente, se presentan los resultados obtenidos en los experimentos de clasificación a partir de las expresiones faciales y voz utilizando técnicas de aprendizaje profundo, así como los resultados de los clasificadores híbridos.

4.1. Conjuntos de datos

Con la finalidad de realizar pruebas al método propuesto para el reconocimiento bimodal de emociones, se realizaron pruebas experimentales con diferentes bases de datos. De estas bases, se destaca nuestra base de datos propuesta UTeMo, la cual es construida especialmente para desarrollar este proyecto de tesis, razón por la cual se describe a detalle a continuación.

4.1.1. Base de datos UTeMo

La base de datos UTeMo está fundamentada en la metodología propuesta en la Sección 3.2. En ésta, cada actor tiene que actuar la emoción a desarrollar tres veces, usando la frase indicada (ver Cuadros 3.1 y 3.2). Se obtuvieron así un total de 1,866

muestras donde se grabaron 294 muestras para las emociones de sorpresa y tristeza, 270 para la emoción neutral y 252 muestras para el resto de las emociones (felicidad, miedo, ira y asco).

En la primer validación donde un experto se asegura que las muestras tengan una correcta expresión facial y entonación de voz, se eliminaron algunas muestras que no cumplían con las características exigidas, es decir, que la emoción representada no se estaba interpretando de una manera correcta. El resultado fue de 1,801 muestras validadas por el experto que se distribuyen de la siguiente manera: ira con 249 muestras; sorpresa, 279; felicidad, 242; miedo, 244; asco, 244; tristeza, 289; neutro, 254 (ver Cuadro 4.1). Para tener un panorama más amplio de esta validación ver la Sección 3.2.

En la segunda validación que consta de una encuesta, se recogieron un total de 638 respuestas. Respecto a los resultados de la encuesta del juicio categórico, hay 485 coincidencias entre las etiquetas originales y las respuestas de los participantes, lo que representa un 76 % de concordancia. En la Figura 4.1 se observa que solamente una fracción menor del total de las muestras fue etiquetada incorrectamente por los encuestados. Asimismo, se puede observar que hay ciertas emociones que las personas consideran similares (por ejemplo, puede haber confusión entre felicidad y sorpresa o entre tristeza y neutral). La razón de esta confusión se debe a la incertidumbre inherente de la percepción humana, por lo que es de esperar que los juicios de los participantes tengan ambigüedades.

La Figura 4.2 muestra los porcentajes de acuerdo con cada emoción en un gráfico heptagonal. En general, se observa que todas las emociones están equilibradas, lo que implica que no hubieron emociones causantes de una gran confusión en los juicios de los participantes.

Cuadro 4.1: Distribución de las muestras de video grabadas para la base de datos propuesta

Emocion	**Número de Muestras originales**	**Número de muestras descartadas**	**Número de muestras aprobadas**
Ira	252	3	249
Sorpresa	294	15	279
Felicidad	252	10	242
Miedo	252	8	244
Asco	252	8	244
Tristeza	294	5	289
Neutro	270	16	254
Totales	1866	65	1801

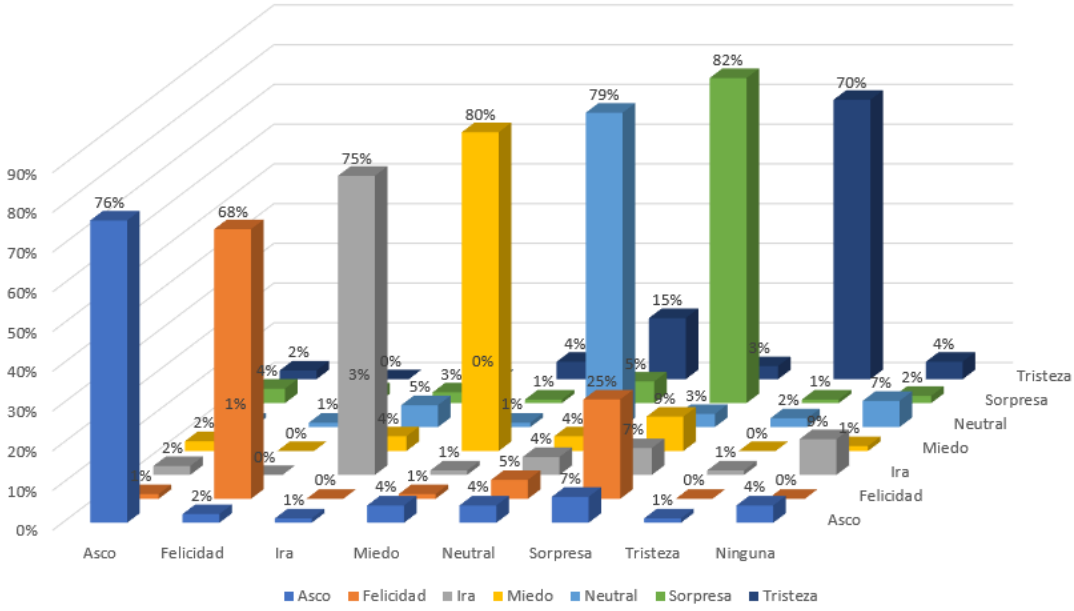


Figura 4.1: Distribución de las respuestas del juicio categórico

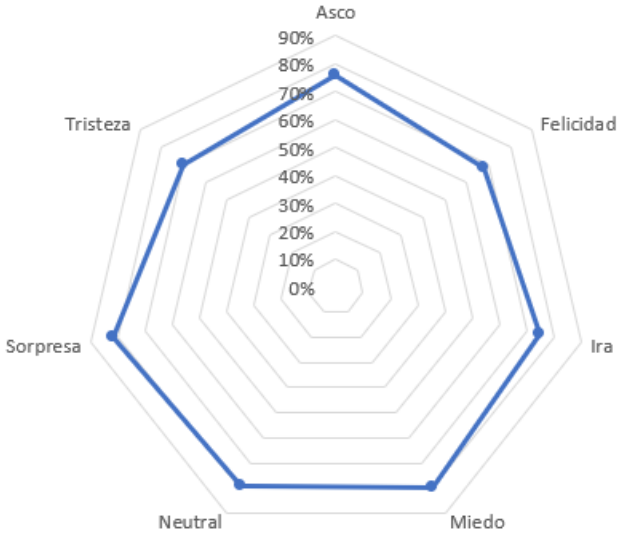


Figura 4.2: Concordancia del juicio categórico por emoción

La distribución de la intensidad para todas las muestras se aprecia en la Figura 4.3. En ésta, se ve que casi la mitad de las muestras son clasificadas como de intensidad media, mientras que la relación entre las muestras correspondientes a las intensidades débil y fuerte es casi la unidad. Este comportamiento es bueno, ya que implica que, a pesar de la incertidumbre de los juicios de los participantes, las emociones que los actores pretendían interpretar sí corresponden en su mayoría con las emociones en cuestión.

Si desglosamos la Figura 4.3 por emoción, se obtiene la Figura 4.4. De esta manera, se puede observar que emociones como la tristeza, el miedo y la sorpresa se perciben con intensidades débiles. Por otro lado, emociones como la felicidad, la ira y el asco se perciben con intensidades fuertes; por último la emoción neutra se percibe con intensidad media. En general, y a excepción de tristeza la mayoría de emociones se percibe con intensidad media por los encuestados.

Con base en los resultados anteriores, se ha calculado el factor Kappa de Fleiss para los juicios categóricos; como se observa en el Cuadro 4.2, se ha obtenido un rendimiento adecuado. En ambos criterios se puede observar que emociones como el asco, la felicidad, la ira y la tristeza son consideradas como regulares de acuerdo con el criterio de [Fleiss et al., 1981] y moderadas de acuerdo con el criterio de [Livingstone and Russo, 2018]; mientras que emociones como el miedo, la sorpresa y neutral son consideradas como buenas según Fleiss y sustanciales según Livingstone y Russo. De manera general, nuestra base de datos UTeMo puede considerarse en ambos criterios como regular, sgún Fleiss y moderada, según Livingstone y Russo. Es importante tener en cuenta que estos valores dependen del juicio subjetivo de los encuestados, por lo que estos valores podrían variar dependiendo de la población a la que se aplique la encuesta.

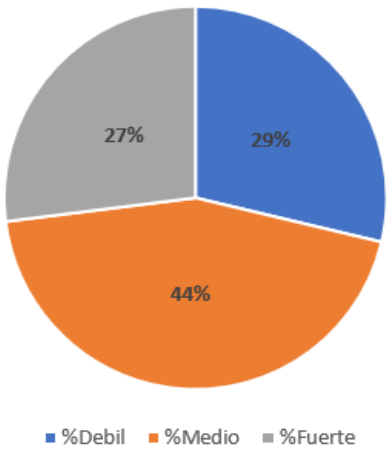


Figura 4.3: Distribución general del juicio de la intensidad

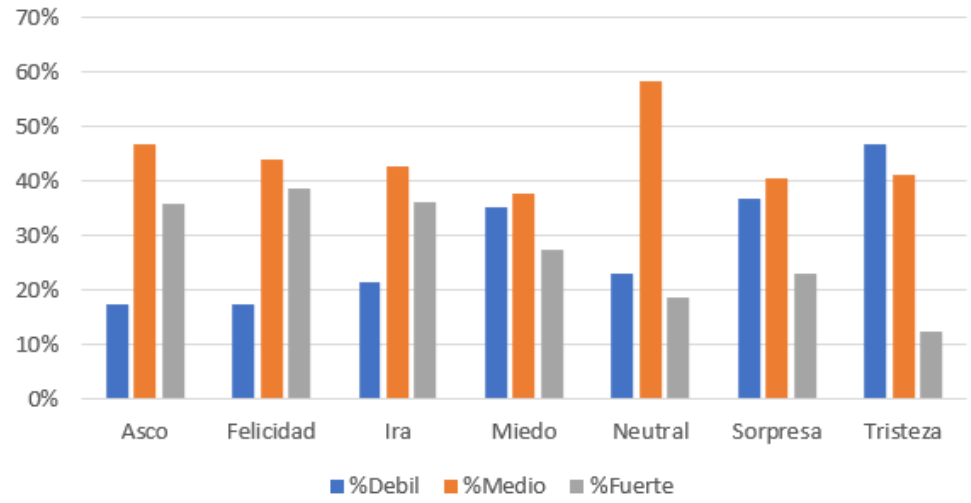


Figura 4.4: Distribución por emoción del juicio de la intensidad

Cuadro 4.2: Factor Kappa (K) de Fleis por emoción y para todas las emociones

Emoción	K	[Fleiss et al., 1981]				[Livingstone and Russo, 2018]					
		< 0.4	0.4 - 0.6	0.61 - 0.75	> 0.75	0	0 - 0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8 - 1
		Malo	Regular	Bueno	Excelente	Pobre	Poco	Razonable	Moderado	Sustancial	Perfecto
Asco	0.57		x						x		
Felicidad	0.58		x						x		
Ira	0.57		x						x		
Miedo	0.64			x						x	
Neutral	0.62			x						x	
Sorpresa	0.66			x						x	
Tristeza	0.51		x						x		
Todas	0.54		x						x		

4.1.2. Preprocesamiento para UTeMo

En el preprocesamiento de audio se utiliza una herramienta de programación para automatizar el proceso que realiza la herramienta Audacity de manera manual. Por sus características, se decidió ocupar la biblioteca `Pydub` de `Python`, la cual también permite trabajar con audio. Internamente, esta herramienta opera con segmentos de audio de $10ms$ de duración, los cuales fueron intencionalmente definidos de este tamaño para los objetivos y metas que quiere alcanzar esta aplicación. El preprocesamiento de audio incluye lo siguiente.

1. Eliminar información no relevante presente en los silencios iniciales y finales de cada muestra de voz.

2. Amplificar la intensidad de voz en 10dB.

3. Minimizar el ruido ambiental presente en las grabaciones mediante el método de muestreo de frecuencias.

El Cuadro 4.3 presenta los resultados obtenidos al aplicar este preprocesamiento respecto a la elimnación de información no deseada, donde se observa una reducción considerable de tiempo. Lo anterior indica que el preprocesameinto aplicado es efectivo y beneficiará al coste computacional al momento de realizar los experimentos posteriores.

Por otro lado, a las imágenes extraídas por cada muestra de video se les aplicaron las siguientes técnicas como parte del preprocesamiento.

1. Mejora de la imagen mediante la ecualización CLAHE

2. Detección de rostros mediante el algoritmo Viola-Jones

Cuadro 4.3: Resumen de duración de las muestras de audio antes y después de ser preprocesadas

Emoción	**Número de muestras de audio**	**Duración total antes del preprocesamiento (hh:mm:ss)**	**Duración promedio de una muestra antes del preprocesamiento**	**Duración total después del preprocesamiento (mm:ss)**	**Duración promedio de una muestra después del preprocesamiento**
Ira	249	00:11:57	2.87s	05:07	1.23s
Sorpresa	279	00:15:34	3.34s	06:59	1.50s
Felicidad	242	00:12:06	3.00s	06:18	1.56s
Miedo	244	00:13:39	3.35s	06:18	1.54s
Asco	244	00:13:27	3.30s	06:43	1.65s
Tristeza	289	00:20:03	4.16s	12:49	2.66s
Neutral	254	00:13.19	3.14s	09:19	2.20s
Totales	1801	01:40:05	2.86s	53:33	1.76s

4.1.3. Otras bases de datos

Para comparar los resultados del método de reconocimiento híbrido de emociones a partir del análisis de voz y de las expresiones faciales propuesto en esta tesis, se utilizaron dos bases de datos adicionales.

RAVDESS: Es una base de datos multimodal (audiovisual) actuada para reconocimiento de emociones en el habla y el canto (en inglés americano). Las muestras se realizaron por 24 actores profesionales cuyo léxico y vocalización es similar en un acento neutro norteamericano. Las emociones consideradas son: calma, felicidad, tristeza, ira, miedo, sorpresa y asco. Cada expresión se produce en dos niveles de intensidad emocional, con una expresión neutral adicional [Livingstone and Russo, 2018]. La distribución ocupada de las muestras de RAVDESS se muestra en el Cuadro 4.4

SAVEE: Esta base de datos actuada fue grabada como requisito previo para el desarrollo de un sistema de reconocimiento automático de emociones. Las grabaciones de la base fueron interpretadas por 4 actores masculinos usando 7 emociones diferentes (felicidad, tristeza, ira, miedo, neutral, sorpresa y asco). Los datos se grabaron en un laboratorio de medios visuales con equipos audiovisuales de alta calidad. La distribución de clases se muestra en el Cuadro 4.5. Cabe mencionar que, una característica interesante de esta base de datos es que utilizan marcas en el rostro de los actores para mejorar al reconocimiento de expresiones a partir de las expresiones faciales y fue grabada en inglés británico [Haq and Jackson, 2010].

Cuadro 4.4: Distribución RAVDESS

Emoción	**Entrenamiento**	**Prueba**	**Total**
Asco	153	39	192
Calma	153	39	192
Felicidad	153	39	192
Ira	153	39	192
Miedo	153	39	192
Neutral	76	20	96
Sorpresa	153	39	192
Tristeza	153	39	192
Totales	1147	293	1440

Cuadro 4.5: Distribución SAVEE

Emoción	Entrenamiento	Prueba	Total
Asco	47	13	60
Felicidad	47	13	60
Ira	47	13	60
Miedo	47	13	60
Neutral	95	25	120
Sorpresa	47	13	60
Tristeza	47	13	60
Totales	377	103	480

4.2. Resultados de clasificación de emociones con métodos tradicionales

Con el objetivo de validar la factibilidad de la base de datos propuesta UTe-Mo, se realizan experimentos de clasificación utilizando métodos tradicionales de extracción de características y clasificación. Para estos experimentos se aplica una validación cruzada estratificada, utilizando 10 particiones del conjunto de datos de entrenamiento. El conjunto total de datos se divide en 80 % para el entrenamiento y 20 % para la prueba. La clasificación realizada se obtiene a partir de diferentes subconjuntos de características provenientes de la voz y de las expresiones faciales.

4.2.1. Clasificación de emociones utilizando expresiones faciales

Para este primer experimento se utiliza una imagen extraída en escala de grises por muestra. Cabe mencionar que, para los experimentos con métodos tradicionales, se ocupó solo la imagen central de cada muestra de video. Estas imágenes tienen una dimensión de 256×256 píxeles, las cuales son aplanadas en un solo vector de 65,536 elementos por imagen. El experimento se realiza entonces para encontrar las características que logran capturar la información (limitada) de las emociones, ya que en esta fase únicamente interesa validar la base de datos construida para tareas de clasificación. A manera de resumen (Sección 3.3.1), el Cuadro 4.6 muestra de forma clara la distribución de las características extraídas con base en las imágenes que corresponden a las expresiones faciales.

En el Cuadro 4.7, se observa que el mejor resultado de clasificación se obtiene ocupando las características HOG transformadas y reducidas con PCA y también

Cuadro 4.6: Distribución por muestra de las características de imagen.

Característica	Núm. de características
Histograma de Gradientes Orientados (HOG)	800
Patrones Binarios Locales (LBP)	433
Variante Uniform de los LBP	900
Total	2133

Cuadro 4.7: Resultados de exactitud de clasificación usando distintos subconjuntos de características extraídas de las expresiones faciales

Grupo de característica	Núm de características	NB	KNN	SVM
Histograma de Gradientes Orientados (HOG)	800	47.74 %	77.09 %	87.09 %
Patrones Binarios Locales (LBP)	433	**53.54 %**	77.02 %	72.90 %
LBP Uniform	900	41.90 %	78.06 %	71.93 %
HOG-PCA (95)	320	51.93 %	**80.64 %**	**88.06 %**
LBP y HOG	1233	**53.54 %**	77.74 %	82.90 %
LBP Uniform y HOG	1700	50.96 %	78.06 %	72.90 %

usando una SVM con kernel Gaussiano. El número 95 que aparece en la fila 5, columna 1 (*HOG-PCA (95)*) hace referencia a que las 320 características que se utilizan describen el 95 % de la varianza de todas las características originales HOG. Este valor se obtuvo realizando diferentes experimentos con diferentes valores de varianza. De igual forma, se aprecia que HOG es el grupo de características que aporta más información para la discriminación de emociones en imágenes, debido a que el método involucra la frecuencia de las orientaciones o direcciones de los píxeles en una vecindad.

4.2.2. Clasificación de emociones utilizando características de voz

La extracción de las características de voz se realiza a partir de las muestras de audio preprocesadas. En total, se obtienen 98 características prosódicas y espectrales (ver Sección 3.3.2), cuya distribución se muestra en el Cuadro 4.8.

Para la clasificación de emociones, se realizan los experimentos de clasificación normalizando las características de manera estándar (media cero y desviación estándar unitaria). Para ello, se utilizan los clasificadores KNN, SVM y NB. Para los clasificadores KNN y SVM, los parámetros adecuados en cada experimento se encuentran mediante una búsqueda de malla. En el caso de KNN se hace una búsqueda

Cuadro 4.8: Distribución por muestra de las características de audio.

Tipo	Característica	Número de características
Prosódicas	Radio de cruce por cero	1
	Centroide	1
	Energía	1
Espectrales	Coeficientes Cepstrales de Mel (MFCC)	26
	Primer derivada MFCC	26
	Segunda derivada MFCC	26
	Coeficientes de Predicción Lineal	17
	Total	98

con $k = 1, 2, ..., 15$; mientras que para SVM se prueban tres kernels diferentes: lineal, polinomial y Gaussiano. Consecuentemente, la búsqueda de malla se realiza variando los parámetors $\gamma = 0{,}001, 0{,}002..,0{,}009, 0{,}015, 0{,}02, 0{,}5, 0{,}9$ y $C = 1, 2, 3, ..., 5$.

En el Cuadro 4.9, se muestran los diferentes experimentos planteados. Es necesario mencionar que el *Grupo Mel* incluye las características de los MFCC, así como la primer y segunda derivada de éstos. También se realizaron experimentos reduciendo las características con PCA sin éxito, por lo cual no se incluyeron en el Cuadro 4.9. De igual forma, se aprecia que el valor más alto de exactitud de clasificación para características de voz se obtiene combinando todas las características (espectrales y prosódicas) y utilizando como clasificador una SVM con kernel Gaussiano. Se puede apreciar que, valores cercanos a este son obtenidos utilizando solamente características espectrales, en especial los relacionados con MFCC. En particular, se puede observar como el subconjunto de características denominado Grupo Mel logra potenciar al máximo el rendimiento del clasificador sencillo KNN, mostrando

Cuadro 4.9: Resultados de exactitud de clasificación usando distintos subconjuntos de características extraídas de muestras de voz

Grupo de característica	Núm de características	NB	KNN	SVM
Radio de cruce por cero	1	24.50 %	20.64 %	26.45 %
Centroide	1	24.19 %	23.54 %	24.83 %
Energía	1	30.32 %	29.03 %	32.35 %
Coeficientes Cepstrales de Mel (MFCC)	26	56.12 %	85.61 %	83.87 %
Primer derivada MFCC	26	48.06 %	71.61 %	76.12 %
Segunda derivada MFCC	26	42.25 %	58.38 %	61.93 %
Grupo Mel	78	56.77 %	**86.77 %**	87.41 %
Coeficientes de Predicción Lineal	17	27.74 %	43.24 %	46.45 %
Características prosódicas	3	32.50 %	40.00 %	34.51 %
Características espectrales	95	**57.41 %**	85.80 %	88.70 %
Todas las características	98	57.09 %	85.48 %	**89.35 %**

la relevancia de estas características para el reconocimiento de emociones.

4.3. Resultados de clasificación de emociones con métodos de aprendizaje profundo

Con el objetivo de implementar un método híbrido para la clasificación de emociones mediante métodos de aprendizaje profundo, se realizaron diferentes experimentos, primero trabajando con las expresiones faciales y con la voz por separado para posteriormente unir estos clasificadores mediante una regla de decisión (ver Sección 3.3). Cabe mencionar que se realizó una validación cruzada (CV por sus siglas en inglés) de 10 iteraciones con el conjunto de entrenamiento para obtener una estimación de generalización de cada modelo obtenido (voz o expresión facial). Para esto, se particionó el conjunto de entrenamiento en 10 subconjuntos, entrenando con 9 subconjuntos y validando con el restante. Para cada combinación, se obtuvo su respectiva medición de exactitud de clasificación y de estos 10 valores se calcularon la media y desviación estándar de dicha medida de rendimiento. Los resultados obtenidos de este proceso se reportan en la segunda columna de los Cuadros 4.11, 4.13, 4.15 y 4.17

4.3.1. Clasificación de emociones con expresiones faciales

En el caso del reconocimiento de emociones a través de expresiones faciales, se realizaron diferentes experimentos, los cuales se presentan en esta sección. Primero, se presentan los resultados obtenidos al utilizar imágenes de rostros (expresiones faciales) como fuentes de información. Segundo, se presentan los resultados obtenidos al utilizar el calculo de flujo óptico. Posteriormente, se realiza una comparación entre arquitecturas, con la finalidad de conocer si es viable el uso de flujo óptico en la clasificación de emociones y también permite identificar qué arquitectura es más conveniente. Finalmente, se busca la mejor configuración de hiperparámetros para lograr un mejor desempeño en las bases de datos utilizadas.

En cuanto a la comparación de exactitud de clasificación usando flujo óptico (siete imágenes) y las expresiones faciales no procesadas (ocho imágenes), los experimentos realizados se muestran en el Cuadro 4.10. En la primera columna, la etiqueta *SecImagenes* corresponde a las ocho imágenes sin procesar; por otro lado, *Franerback 1* y *Franerback 2*, corresponden a las imágenes generadas de flujo óptico con el método Franerback con los métodos de selección de pares 1 y 2, descritos en la

Cuadro 4.10: Comparación del desempeño en exactitud de clasificación de diferentes modelos de red utilizando flujo óptico e imágenes no procesadas

	MobileNet	**ResNet50**	**VGG16**	**3DCNN**
SecImagenes	98.37 %	97.23 %	16.71 %	50.31 %
Franerback 1	64.75 %	82.81 %	15.56 %	80.60 %
Franerback 2	66.70 %	88.75 %	17.51 %	81.03 %

sección 3.3.1. Cabe mencionar que, estos experimentos se realizaron con la base de datos SAVEE y extrayendo ocho imágenes por muestra, posteriormente se redujeron a cinco imágenes por muestra para comparar cuando se eliminan imágenes al inicio y al final de la muestra, ya que generalmente no presentan una emoción (ver Sección 3.3.1). Los hiperparámetros utilizados para estos experimentos son fijos debido a que el objetivo es conocer qué fuente de información y arquitectura son mejores para esta tarea. La configuración utilizada en los experimentos es la siguiente.

- Número de épocas de entrenamiento: 50.
- Tamaño de lote: 50.
- Función de pérdida: Entropía cruzada categórica.
- Optimizador: Adam con tasa de aprendizaje de 0.001.
- Inicializador: ImageNet

De acuerdo con estos experimentos, se llegó a la conclusión de que la forma adecuada de hacer reconocimiento de emociones es con las imágenes de las expresiones faciales no procesadas, sin necesidad de calcular el flujo óptico. Ésto es, no sólo porque los experimentos con esta fuente de infomación (ver Cuadro 4.10) presentan mejor desempeño de exactitud de clasificación, sino también porque al no hacer el cálculo de flujo óptico supone un ahorro de costo computacional. Asimismo, se aprecia que la arquitectura más viable es la MobileNet, seguida de la ResNet50. Cabe mencionar que la MobileNet es más ligera (en términos de número total de pesos sinápticos) que la ResNet50, por lo que utilizar esta red también supone un beneficio en el costo (tiempo) computacional.

Con base en los resultados anteriores, se procede a realizar una búsqueda de malla para encontrar los mejores hiperparámetros de la arquitectura MobileNet usando las distintas bases de datos ocupadas (SAVEE, RAVDESS y UTeMo), como se menciona en la Sección 3. Los resultados obtenidos se muestran en el Cuadro 4.11 y los hiperparámetros para cada base de datos se presentan en el Cuadro 4.12.

Cuadro 4.11: Resultados de medidas de rendimiento utilizando la arquitectura MobileNet con expresiones faciales. La abreviación CV significa validación cruzada y CT significa conjunto de prueba

Base de datos	Exactitud (CV)	Exactitud (CT)	F1-Score (CT)	Precisión (CT)
UTeMo	$93.44 \pm 4.95\,\%$	$95.96\,\%$	$95.85\,\%$	$95.85\,\%$
SAVEE	$99.25 \pm 0.63\,\%$	$95.41\,\%$	$94.71\,\%$	$94.85\,\%$
RAVDESS	$91.72 \pm 1.39\,\%$	$88.97\,\%$	$88.62\,\%$	$88.37\,\%$

Cuadro 4.12: Hiperparámetros encontrados para el reconocimiento de emociones utilizando la arquitectura MobileNet con expresiones faciales

Base de datos	# épocas	Optimizador	Tamaño de lote	Inicializador de pesos sinápticos
UTeMo	100	Adam	50	ImageNet
SAVEE	50	SGD	50	ImageNet
RAVDESS	100	Adam	50	ImageNet

En el Cuadro 4.11 se puede apreciar que, los experimentos con UTeMo y SAVEE tienen desempeños similares (en conjunto de prueba), aún cuando UTeMo no utiliza las marcas faciales en los rostros de actores. Esta similitud se debe a que UTeMo cuenta con un número mucho mayor de muestras, lo cual beneficia al reconocimiento de emociones a partir de expresiones faciales. De igual forma, con estos experimentos se comprueba lo que se menciona en la literatura [Darwin, 1872, Ekman, 1992] acerca que las expresiones faciales son universales para representar emociones, ya que hay un buen desempeño utilizando todas las bases de datos.

4.3.2. Clasificación de emociones con muestras de voz

El objetivo de estos experimentos es implementar un método funcional para realizar reconocimiento de emociones mediante información de la voz. Para éstos, se utilizan redes neuronales convolucionales como clasificadores principales. En esta sección, se presentan los resultados obtenidos al utilizar la voz como principal fuente de información. Primeramente, se realizan experimentos para encontrar un conjunto de datos aumentado adecuado para trabajar con la arquitectura 2D FSER (ver Sección 2.5.2). Posteriormente, se busca la mejor configuración de hiperparámetros para lograr un mejor desempeño al utilizar espectrogramas y cocleogramas. Finalmente se muestran los resultados obtenidos al trabajar con las arquitecturas 1D (igualmente descritas en la Sección 2.5.2).

El primer grupo de experimentos consiste en realizar una búsqueda de malla gruesa variando el ángulo de rotación y el factor de acercamiento/alejamiento aplicado al conjunto empleado. Esto, con el fin de encontrar el ángulo y factor más

apropiados para la clasificación. Aun cuando en la práctica no se van a introducir representaciones espectrales rotadas o acercadas/alejadas a la red neuronal, este proceso es útil, ya que ayuda a la red a generalizar la relación espacial entre frecuencias e intensidades que existen en cada determinada emoción.

Los experimentos de clasificación para encontrar un aumento de datos apropiado se realizan con la arquitectura FSER (ver Figura 2.25) utilizando los siguientes hiperparámetros.

- Optimizador: SGD con una constante de aprendizaje de 0.001.
- Tamaño de lote: 64.
- Entrenamiento: 400 épocas.
- Inicializador: Glorot Normal.

Al realizar estos experimentos, los mejores resultados se obtuvieron con un factor de acercamiento/alejamiento de 0.05 y un ángulo de rotación de 5°, por lo que se procede a trabajar con este conjunto de datos aumentado.

Una vez definido el conjunto aumentado se realiza una búsqueda de malla adicional para encontrar los hiperparámetros adecuados de la arquitectura FSER aplicada a UTeMo, SAVEE y RAVDESS, tanto para los espectrogramas de Mel como para los cocleogramas.

Se observa en el Cuadro 4.13 que los resultados obtenidos con SAVEE y RAVDESS difieren con respecto a los obtenidos por [Dossou and Gbenou, 2021]. En este cuadro se ve una gran diferencia de desempeño en función de la base de datos, lo cual se debe al número de muestras en cada base y a los fonemas que se ocupan en cada idioma. Lo anterior confirma que, las emociones mediante el uso de la voz no son universales y varía la forma de interpretación de acuerdo al lenguaje. Los hiperparámetros encontrados se presentan en el Cuadro 4.14.

Cuadro 4.13: Resultados de medidas de rendimiento utilizando la arquitectura FSER con espectrogramas de voz. La abreviación CV significa validación cruzada y CT significa conjunto de prueba

Base de datos	**Exactitud (CV)**	**Exactitud (CT)**	**F1-Score (CT)**	**Precisión (CT)**
UTeMo	$99.03 \pm 0.09\,\%$	$93.44\,\%$	$93.42\,\%$	$93.42\,\%$
SAVEE	$87.75 \pm 2.01\,\%$	$61.81\,\%$	$58.28\,\%$	$63.00\,\%$
RAVDESS	$62.16 \pm 2.26\,\%$	$63.99\,\%$	$62.62\,\%$	$63.75\,\%$

Por otro lado, en el Cuadro 4.15 se observa que, para las bases de datos en inglés (SAVEE y RAVDESS) los cocleogramas no son útiles para hacer reconocimiento de emociones mediante la voz. Además, durante el entrenamiento se observa que en el caso de RAVDESS, el modelo no pudo extraer las características adecuadas; y en el caso de SAVEE hubo un sobreentrenamiento. Sin embargo, se resalta que con la base de datos UTeMo se logran muy buenos resultados (en conjunto de prueba), incluso superiores a los obtenidos con espectrogramas, lo cual indica que esta representación espectral es viable para hacer reconocimiento de emociones a partir de la voz, al menos para el español de México. Los hiperparámetros encontrados se muestran en el Cuadro 4.16.

Finalmente, respecto a los experimentos realizados con arquitecturas 1D (ver Figura 2.26 descritos en la Sección 3.3.2), se realizó una búsqueda de malla para determinar la mejor configuración. En total, se utilizaron 182 características, las cuales se describieron en la Sección 3.3.2. Los resultados correspondientes se muestran en el Cuadro 4.17. En este caso, estos resultados son más consistentes con la literatura para reconocimiento de emociones a partir de la voz [Middya et al., 2022, Issa et al., 2020]. De igual forma se aprecia que, el experimento realizado con UTeMo tiene un mejor desempeño respecto a las otras bases de datos. Los hiperparámetros

Cuadro 4.14: Hiperparámetros encontrados para la arquitectura FSER con espectrogramas de voz

Base de datos	**# épocas**	**Optimizador**	**Tamaño de lote**	**Inicializador de pesos sinápticos**
UTeMo	400	SGD	32	He Normal
SAVEE	100	SGD	32	Glorot Normal
RAVDESS	200	Adam	64	He Normal

Cuadro 4.15: Resultados de medidas de rendimiento utilizando la arquitectura FSER con cocleogramas de voz. Los resultados marcados con "-" no se pudieron calcular debido a los resultados en la matriz de confusión. La abreviación CV significa validación cruzada y CT significa conjunto de prueba

Base de datos	**Exactitud (CV)**	**Exactitud (CT)**	**F1-Score (CT)**	**Precisión (CT)**
UTeMo	99.43 ± 0.09	95.08 %	95.28 %	95.42 %
SAVEE	96.05 ± 0.68	23.63 %	-	-
RAVDESS	24.84 ± 2.69	23.64 %	-	-

Cuadro 4.16: Hiperparámetros encontrados para la arquitectura FSER con cocleogramas de voz

Base de datos	**# épocas**	**Optimizador**	**Tamaño de lote**	**Inicializador de pesos sinápticos**
UTeMo	100	Adam	32	Glorot Normal
SAVEE	400	SGD	64	Glorot Normal
RAVDESS	100	Adam	64	He Normal

Cuadro 4.17: Resultados de medidas de rendimiento utilizando la arquitectura 1D con características de voz. La abreviación CV significa validación cruzada y CT significa conjunto de prueba

Base de datos	Exactitud (CV)	Exactitud (CT)	F1-Score (CT)	Precisión (CT)
UTeMo	$74.47 \pm 2.33\,\%$	$81.96\,\%$	$81.71\,\%$	$82.28\,\%$
SAVEE	$56.03 \pm 7.92\,\%$	$63.63\,\%$	$61.28\,\%$	$67.57\,\%$
RAVDESS	$65.58 \pm 2.35\,\%$	$66.00\,\%$	$65.62\,\%$	$65.87\,\%$

Cuadro 4.18: Hiperparámetros encontrados para la arquitectura 1D con características de voz

Base de datos	# épocas	Optimizador	Tamaño de lote	Inicializador de pesos sinápticos
UTeMo	200	Adam	16	Glorot Normal
SAVEE	300	Adam	64	He Normal
RAVDESS	300	SGD	16	He Normal

encontrados se presentan en el Cuadro 4.18.

De manera general, los experimentos obtenidos con UTeMo muestran un mejor desempeño con respecto al resto de bases de datos empleadas. Una de las razones principales de este comportamiento se debe a que UTeMo fue construida con un número superior de actores, siendo superada en la literatura sólo por RAVDESS. Otra ventaja de UTeMo es que para su construcción se consideraron frases específicas para cada emoción. De igual forma el idioma influye en los resultados, por lo cual el español mexicano tiene cierta ventaja sobre el inglés para el reconocimiento de emociones, ya que utiliza un número inferior de fonemas, lo cual influye en la entonación de la voz al expresar emociones [Adelman et al., 2018].

4.3.3. Clasificación de emociones de manera híbrida

En esta sección se muestran los resultados obtenidos al combinar la voz y expresiones faciales como fuentes de información, es decir, los resultados del método híbrido propuesto para el reconocimiento de emociones.

Con base en los resultados obtenidos, de acuerdo con lo descrito en la Sección 3.3.3, se procede a buscar algún valor α apropiado en las diferentes combinaciones utilizando los diversos modelos obtenidos en las secciones anteriores, por lo que se realiza una búsqueda variando este factor. Dicha búsqueda consiste en variar el factor de 0.1 a 0.9 con intervalos de 0.1. Se presentan los mejores resultados del parámetro α de esta búsqueda en sus diferentes combinaciones de modelos de voz en los Cuadros 4.19, 4.20 y 4.21. En la práctica, un valor de α entre 0.1 y 0.4 indica que se le da

preferencia a la decisión del clasificador obtenido con voz, mientras que un valor entre 0.6 y 0.9 da mayor importancia a la decisión del clasificador obtenido con expresiones faciales.

En el Cuadro 4.19 se aprecia que el modelo propuesto para UTeMo mantiene un α equilibrado, mientras que los modelos propuestos para SAVEE y RAVDESS se inclinan por la decisión basada en expresiones faciales, ya que entre mayor sea α, mayor peso tiene esta decisión.

Respecto a los resultados de los experimentos del Cuadro 4.20, vemos que se presenta la mayor exactitud de clasificación para el modelo usado en UTeMo y un α equilibrado; mientras que, para los modelos propuestos para SAVEE y RAVDESS la decisión basada en expresiones faciales tiene mucho mayor peso que la de voz, debido al pésimo desempeño obtenido utilizando cocleogramas.

Además, en el Cuadro 4.21 podemos observar que los modelos propuestos para SAVEE y RAVDESS logran un mejor desempeño comparado con los modelos en los otros grupos de experimentos (ver Cuadros 4.19 y 4.20). De igual forma, los modelos

Cuadro 4.19: Resultados obtenidos con información híbrida utilizando MobileNet y FSER con espectrogramas

Base de datos	α	**Exactitud**	**F1-Score**	**Precisión**
UTeMo	0.5	98.89 %	98.92 %	98.92 %
SAVEE	0.6	95.80 %	95.25 %	95.34 %
RAVDESS	0.6	90.06 %	89.70 %	89.79 %

Cuadro 4.20: Resultados obtenidos con información híbrida utilizando MobileNet y FSER con cocleogramas

Base de datos	α	**Exactitud**	**F1-Score**	**Precisión**
UTeMo	0.5	99.26 %	99.25 %	99.22 %
SAVEE	0.9	95.41 %	94.79 %	94.85 %
RAVDESS	0.9	88.97 %	88.37 %	88.64 %

Cuadro 4.21: Resultados obtenidos con información híbrida utilizando MobileNet y 1D con características de voz

Base de datos	α	**Exactitud**	**F1-Score**	**Precisión**
UTeMo	0.5	97.30 %	97.25 %	97.28 %
SAVEE	0.6	96.18 %	95.69 %	95.81 %
RAVDESS	0.5	92.78 %	92.67 %	93.05 %

propuestos para RAVDESS y UTeMo mantienen un α equilibrado, mientras que el modelo propuesto para SAVEE se inclina un poco más hacia las expresiones faciales.

De manera general, se aprecia que, al combinar las decisiones de dos clasificadores se obtiene un mejor clasificador final, lo cual indica que el reconocimiento híbrido o multimodal de emociones es mejor que el reconocimiento unimodal. Además, al combinar la decisión de los clasificadores de expresiones faciales y voz, el rendimiento del modelo con expresiones faciales amortigua el efecto del mal desempeño del clasificador de voz para las bases de datos en inglés.

Con base en los mejores resultados obtenidos, en el Cuadro 4.22 se presenta una comparación con los modelos más recientes propuestos en la literatura enfocados en el reconocimiento bimodal a partir de las expresiones faciales y el análisis de voz. Se aprecia que nuestros modelos son bastante competitivos con respecto a los presentados en la literatura e incluso mejores que algunos de éstos.

Finalmente, en el Cuadro 4.23 podemos apreciar el tiempo de ejecución de las partes que conforman nuestra mejor propuesta usando la base de datos propuesta UTeMo, la cual considera el uso de la representación espectral de cocleogramas para el reconocimiento de voz y la arquitectura MobileNet para las expresiones faciales. Estos tiempos de ejecución se tomaron utilizando cómputo secuencial y cómputo paralelo, con el equipo descrito en la Sección 3.1. En el Cuadro 4.23 podemos ver las diferentes aceleraciones que se obtuvieron durante el entrenamiento, donde la aceleración en el entrenamiento de la red FSER con cocleogramas es de casi seis veces más, esto se debe a que con esta fuente de información se procesa una mayor cantidad de imágenes que con expresiones faciales mediante la arquitectura MobileNet.

Cuadro 4.22: Comparación de nuestra propuesta con modelos recientes de la literatura

Base de datos	**Método**	**Arquitectura**	**CNN 1D/2D**	**Exactitud de clasificación**
UTeMo	Nuestra propuesta	MobileNet + Cocleogramas	2D + 2D	99.26 %
SAVEE	Nuestra propuesta	MobileNet + 1D	2D + 1D	96.18 %
SAVEE	[Middya et al., 2022]	1D básica	1D + 1D	99.00 %
SAVEE	[Avots et al., 2019]	AlexNet + SVM	2D + Tradicional	94.33 %
RAVDESS	Nuestra propuesta	MobileNet + 1D	2D + 1D	92.78 %
RAVDESS	[Middya et al., 2022]	1D básica	1D + 1D	86.00 %

Cuadro 4.23: Aceleración obtenida durante el entrenamiento de los diferentes modelos

Arquitectura (representación)	**Secuencial**	**Paralelo**	**Aceleración**
FSER (cocleogramas)	10.30h	1.73h	5.95x
MobileNet (rostros)	4.24h	2.03h	2.10x

Capítulo 5

Conclusiones y trabajo a futuro

En esta tesis se presenta a UTeMo, una nueva base de datos audiovisual que ha sido validada tanto cualitativa (mediante la opinión de un experto y mediante encuestas), como cuantitativamente (mediante el uso de técnicas de aprendizaje computacional). Su factibilidad permite que UTeMo sea fiable para tareas de reconocimiento de emociones en el contexto del español mexicano. Además, con UTeMo se fomentará la investigación científica en el reconocimiento de emociones en español de México. De igual forma, en este trabajo se implementa un nuevo modelo de reconocimiento híbrido de emociones a partir del análisis de la voz y de las expresiones faciales, mejorando así el desempeño de clasificación de las emociones básicas mediante técnicas de aprendizaje profundo al combinar las decisiones de ambos clasificadores (voz y expresiones faciales). Lo anterior se logró debido a la metodología llevada a cabo tanto para la construcción de la base de datos como para su validación.

En este trabajo, se ha verificado lo que se menciona en la literatura acerca de la universalidad emocional de las expresiones faciales [Sasaki, 1993, Darwin, 1872], debido a que con esta fuente de información se obtuvieron resultados más consistentes en los diferentes modelos propuestos. En cambio, los resultados de los modelos para reconocer emociones a partir de la voz varían dependiendo de la base de datos empleada (por ejemplo, los modelos 1D beneficiaron más a las bases de datos en inglés SAVEE y RAVDESS); mientras que para UTeMo se logró un mejor desempeño con la arquitectura 2D FSER. Además, se validó experimentalmente que la combinación de un clasificador fuerte (expresiones faciales) con uno débil (cocleogramas) ayuda a lograr un mejor desempeño en general y en este estudio se hace notorio para la base de datos UTeMo.

Cabe mencionar que en este trabajo de tesis el reconocimiento de emociones

híbrido presenta mejores resultados en general, y en particular para la base UTeMo alcanza un 99.26 % de exactitud de clasificación, que sólo haciendo reconocimiento de emociones mediante la voz o expresiones faciales por separado, con un 95.08 % y 95.96 % de exactitud, respectivamente.

Debido a que el análisis de voz se centró en bases de datos en ingles y español de México, a futuro se planea ampliar este estudio incluyendo otras bases de datos en diferentes idiomas, así como otras arquitecturas de red específicas para cada base de datos y otras representaciones espectrales. De esta manera, podría obtenerse un conocimiento más profundo acerca de la importancia del idioma y las consideraciones que hay que tener al momento de hacer reconocimiento de emociones a partir de la voz.

Bibliografía

[Adelman et al., 2018] Adelman, J. S., Estes, Z., and Cossu, M. (2018). Emotional sound symbolism: Languages rapidly signal valence via phonemes. *Cognition*, 175:122–130.

[Aggarwal C., 2018] Aggarwal C., C. (2018). *Neural Networks and Deep Learning*. Springer, New york, USA, 1st edition.

[Aladem and Rawashdeh, 2020] Aladem, M. and Rawashdeh, S. A. (2020). A single-stream segmentation and depth prediction cnn for autonomous driving. *IEEE Intelligent Systems*, 36(4):79–85.

[Albawi et al., 2017] Albawi, S., Mohammed, T. A. M., and Alzawi, S. (2017). *Understanding of a Convolutional Neural Network*. In *IEEE International Conference on Engineering and Technology*, pages 1–6, Antalay, Turkey.

[Alvarado Moya, 2012] Alvarado Moya, J. P. (2012). *Procesamiento y Análisis de Imágenes Digitales [Tesis de Ingeniería]*. Instituto Tecnológico de Costa Rica, Escuela de Ingeniería en Electrónica.

[Alvarez and Guevara, 2009] Alvarez, D. and Guevara, M. (2009). Reconocimiento de expresiones faciales prototipo usando ICA. *Scientia Et Technica*, 1(41):81–86.

[Antoniadis et al., 2021] Antoniadis, P., Pikoulis, I., Filntisis, P. P., and Maragos, P. (2021). An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3645–3651, Online.

[Anvita et al., 2020] Anvita, S., Ashish, K., and Deepak, G. (2020). Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79.

[Avots et al., 2019] Avots, E., Sapiński, T., Bachmann, M., and Kamińska, D. (2019). Audiovisual emotion recognition in wild. *Machine Vision and Applicationss*, 30(5):975 – 985.

[Ayadi et al., 2011] Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(1):572–587.

[Barra-Chicote et al., 2008] Barra-Chicote, R., Montero, J., Macias-Guarasa, J., Lebai Lutfi, S., Lucas-Cuesta, J., Fernández-Martínez, F., D'Haro, L., Hernandez, R., Ferreiros, J., Cordoba, R., and Muñoz, J. (2008). Spanish expressive voices: Corpus for emotion research in spanish. In *6th conference of Language Resources Evaluation*, pages 60–70, Morroco.

[Bhardwaj et al., 2018] Bhardwaj, A., Di, W., and Wei, J. (2018). *Deep Learning Essentials: Your hands-on guide to the fundamentals of deep learning and neural network modeling*. Packt Publishing Ltd.

[Birajdar and Patil, 2020] Birajdar, G. and Patil, M. (2020). Speech/music classification using visual and spectral chromagram features. *Journal of Ambient Intelligence and Humanized Computing*, 11:329–347.

[Bisquerra Alzina, 2009] Bisquerra Alzina, R. (2009). *Psicopedagogía de las Emociones*. Sintesis, Madrid, 1er edición.

[Boersma and Weenink, 2021] Boersma, P. and Weenink, D. (2021). Praat: doing phonetics by computer. [Online]. Available (2021, September): `https://www.fon.hum.uva.nl/praat/`.

[Burkhardt et al., 2005] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, Lissabon, Portugal.

[Calvo et al., 2014] Calvo, R., D'Mello, S., Gratch, J., and Kappas, A. (2014). *Introduction to Affective Computing*, volume 1, pages 1–15. Oxford University Press.

[Canny, 1986] Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698.

[Carleo et al., 2019] Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):1–39.

[Catrillon et al., 2008] Catrillon, W., Alvarez, D., and López, A. (2008). Técnicas de extracción de características en imágenes para el reconocimiento de expresiones faciales. *Scientia Et Technica*, 14(38):7–12.

[Chandrasekar et al., 2014] Chandrasekar, P., Chapaneri, S., and Jayaswal, D. (2014). Automatic speech emotion recognition: A survey. In *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*, pages 341–346, Mumbai, India.

[Chowdary et al., 2021] Chowdary, M. K., Nguyen, T. N., and Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, pages 1–18.

[Cicres, 2009] Cicres, J. (2009). Evaluación del uso de cocleagramas para la identificación del hablante en fonética forense. *Síntesis Tecnológica*, 4(1):37–46.

[Damasio, 1999] Damasio, A. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Mariner Books, London, 1st edition.

[Darwin, 1872] Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. John Murray, United Kingdom, 1st edition.

[Davidson et al., 2018] Davidson, A. S. F., Lapate, R. C., Shackman, A. J., and J, R. (2018). *The nature of emotion : fundamental questions*. Oxford University Press, New York, 2nd edition.

[De Marchi and Mitchell, 2019] De Marchi, L. and Mitchell, L. (2019). *Hands-On Neural Networks: Learn how to build and train your first neural network model using Python*. Packt Publishing Ltd.

[Delgado and Mora, 1998] Delgado, J. M. and Mora, T. (1998). *Manual de Neurociencias*. Editorial Síntesis, Madrid, España, 1st edition.

[Deng et al., 2020] Deng, D., Chen, Z., Zhou, Y., and Shi, B. (2020). Mimamo net: Integrating micro and macro motion for video emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2621–2628.

[Deng et al., 2013] Deng, L., He, X., and Gao, J. (2013). Deep stacking networks for information retrieval. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3153–3157, Vancouver, BC, Canada.

[Developer, 2021] Developer, I. (2021). Arquitecturas de apren-

dizaje profundo. [Online]. Available (2021, Abril): https://developer.ibm.com/es/technologies/deep-learning/articles/cc-machine-learning-deep-learning-architectures/.

[Dino et al., 2020] Dino, H., Abdulrazzaq, M. B., Zeebaree, S., Sallow, A. B., Zebari, R. R., Shukur, H. M., and Haji, L. M. (2020). Facial expression recognition based on hybrid feature extraction techniques with different classifiers. *TEST Engineering and Management*, 83:22319–22329.

[Dino and Abdulrazzaq, 2019] Dino, H. I. and Abdulrazzaq, M. B. (2019). Facial expression classification based on SVM, KNN and MLP classifiers. In *2019 International Conference on Advanced Science and Engineering (ICOASE)*, pages 70–75, Zakho-Duhok, Iraq.

[Dossou and Gbenou, 2021] Dossou, B. F. and Gbenou, Y. K. (2021). FSER: Deep convolutional neural networks for speech emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3533–3538, Virtual.

[Du et al., 2020] Du, G., Long, S., and Yuan, H. (2020). Non-Contact Emotion Recognition Combining Heart Rate and Facial Expression for Interactive Gaming Environments. *IEEE Access*, 8:11896–11906.

[Díaz Salcedo and Higuera Martínez, 2006] Díaz Salcedo, L. N. and Higuera Martínez, I. (2006). Estimación de movimiento en imágenes de resonancia magnética cardiaca. *Tecnura*, 9(18):27–35.

[Ekman, 1992] Ekman, P. (1992). An Argument for Basic Emotions. *American Psychologist*, 6(3):169–200.

[Ekman et al., 1983] Ekman, P., Friesen, W., and Levenson, R. (1983). Autonomic Nervous System Activity Distinguishes Among Emotions. *Science*, 221(4616):1208–1210.

[Elliott et al., 2011] Elliott, R., Bohart, A. C., Watson, J. C., and Greenberg, L. S. (2011). Empathy. *Psychotherapy*, 48(1):43 – 49.

[Fan et al., 2016] Fan, Y., Lu, X., Li, D., and Liu, Y. (2016). *Video-Based emotion recognition using CNN-RNN and C3D hybrid networks*. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450, New York, United States.

[Fernández-Abascal and Cano-Vindel, 1995] Fernández-Abascal, E. and Cano-Vindel, A. (1995). Cognición y emoción. *Manual de Motivación y Emoción*, 10:113–160.

[Fleiss et al., 1981] Fleiss, J. L., Levin, B., Paik, M. C., et al. (1981). The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

[Gong et al., 2019] Gong, C., Lin, F., Zhou, X., and Lu, X. (2019). Amygdala-inspired affective computing: To realize personalized intracranial emotions with accurately observed external emotions. *China Communications*, 16(8):115–129.

[Gonzales C. and Woods E., 2008] Gonzales C., R. and Woods E., R. (2008). *Digital Imagen Processing*. Prentice Hall, Muchigan, USA, 3th edition.

[Gonzáles and Wintz, 1996] Gonzáles, R. and Wintz, P. (1996). *Filtrado de Imágenes*, chapter 3, pages 89–269. Universidad de Tarapacá, Addison Wesley, 1st edition.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Massachusett, USA, 1st edition.

[Graves, 2012] Graves, A. (2012). *Long Short-Term Memory*, chapter 4, pages 37–45. Technical University of Munich, Toronto, Ontario, 1st edition.

[Haq and Jackson, 2010] Haq, S. and Jackson, P. (2010). *Machine Audition: Principles, Algorithms and Systems*, chapter Multimodal Emotion Recognition, pages 398–423. IGI Global, Hershey PA.

[Haykin, 2009] Haykin, S. (2009). *Neural Networks and Learning Machines*. Prentice Hall, New york, USA, 3th edition.

[He et al., 2020] He, Z., Li, Z., Yang, F., Wang, L., Li, J., Zhou, C., and Pan, J. (2020). Advances in multimodal emotion recognition based on brain–computer interfaces. *Brain Sciences*, 10(10):1–29.

[Hilera Gonzales and Martínez Hernando, 1995] Hilera Gonzales, J. R. and Martínez Hernando, V. J. (1995). *Redes neuronales artificiales: fundamentos, modelos y aplicaciones*. Addion-Wesley, Buenos Aires, Argentina, 1st edition.

[Hippe et al., 2014] Hippe, Z. S., Kulikowski, J. L., Mroczek, T., and Wtorek, J. (2014). Human-Computer Systems Interaction: Backgrounds and Applications.

Advances in Intelligent Systems and Computing, 300:51–62.

[Horn and Schunck, 1981] Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, 17(1):185–203.

[Hossain and Muhammad, 2017] Hossain, M. S. and Muhammad, G. (2017). An Emotion Recognition System for Mobile Applications. *IEEE Access*, 5:2281–2287.

[Hossain and Muhammad, 2019] Hossain, M. S. and Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78.

[Howard et al., 2017] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[Issa et al., 2020] Issa, D., Demirci, M. F., and Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:1–10.

[Izard, 1991] Izard, C. E. (1991). *The Psychology of Emotions*. Springer US, New York, US, 1st edition.

[Ji et al., 2021] Ji, Y., Zhang, H., Zhang, Z., and Liu, M. (2021). Cnn-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances. *Information Sciences*, 546:835–857.

[Kass et al., 1988] Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331.

[Khan et al., 2020] Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455 – 5516.

[Kim et al., 2020] Kim, H., Ben-Othman, J., Mokdad, L., and Lim, K. (2020). CONTVERB: Continuous Virtual Emotion Recognition Using Replaceable Barriers for Intelligent Emotion-Based IoT Services and Applications. *IEEE Network*, 34(5):269–275.

[Kim and Provost, 2019] Kim, Y. and Provost, E. M. (2019). ISLA: Temporal Segmentation and Labeling for Audio-Visual Emotion Recognition. *IEEE Transac-*

tions on Affective Computing, 10(2):196–208.

[Koduru et al., 2020] Koduru, A., Valiveti, H. B., and Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology*, 23(1):45–55.

[Kołakowska et al., 2014] Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., and Wrobel, M. R. (2014). *Emotion Recognition and Its Applications*, chapter 5, pages 51–62. Springer, Switzerland, 1st edition.

[Konar and Chakraborty, 2015] Konar, A. and Chakraborty, A. (2015). *Emotion Recognition: A Pattern Analysis Approach*. Wiley, Hoboken, New Jersey, 1st edition.

[Koolagudi and Rao, 2020] Koolagudi, S. G. and Rao, K. S. (2020). Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117.

[Liu et al., 2011] Liu, Y., Sourina, O., and Nguyen, M. K. (2011). Real-Time EEG-Based Emotion Recognition and Its Applications. *Transactions on Computational Science*, 12:256–277.

[Liu et al., 2017] Liu, Z., Wu, M., Cao, W., Chen, L., Xu, J., Zhang, R., Zhou, M., and Mao, J. (2017). A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica*, 4(4):668–676.

[Livingstone and Russo, 2018] Livingstone, S. R. and Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLoS ONE*, 13(5):1–35.

[López et al., 2006] López, J. M., Cearreta, I., Garay, N., López de Ipiña, K., and Beristain, A. (2006). Creación de una base de datos emocional bilingüe y multimodal. *Innovae Visión*, 7:55–64.

[López Gil and Garay Vitoria, 2019] López Gil, J. M. and Garay Vitoria, N. (2019). Emotion recognition in video and audio through the use of artificial intelligence techniques. In *ACM International Conference Proceeding Series*, pages 1–2, Donostia, España.

[Lowhur and Chuah, 2015] Lowhur, A. and Chuah, M. C. (2015). Dense optical flow based emotion recognition classifier. In *2015 IEEE 12th International Conference on Mobile Ad Hoc and Sensor Systems*, pages 573–578, Dallas, TX, USA.

[Matsumoto et al., 2013] Matsumoto, D., Hwang, H. S., López, R. M., and Pérez Nieto, M. A. (2013). Lectura de la expresión facial de las emociones: Investigación básica en la mejora del reconocimiento de emociones. *Ansiedad y estrés*, 19(2-3):121–129.

[Meng et al., 2019] Meng, H., Yan, T., Yuan, F., and Wei, H. (2019). Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE access*, 7:125868–125881.

[Middya et al., 2022] Middya, A. I., Nag, B., and Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems*, 244:108580.

[Moolchandani et al., 2021] Moolchandani, M., Dwivedi, S., Nigam, S., and Gupta, K. (2021). A survey on: Facial emotion recognition and classification. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1677–1686, Erode, India.

[Mora Teruel, 2013] Mora Teruel, F. (2013). ¿Qué Es Una Emoción? *Arbor Ciencia, Pensamiento y Cultura*, 189(759):1–6.

[Mora Teruel and Sanguinetti, 2004] Mora Teruel, F. and Sanguinetti, A. M. (2004). *Diccionario de Neurociencia*. Alianza, España, 1st edition.

[Moya Albiol, 2014] Moya Albiol, L. (2014). *La empatía: Entenderla para entender a los demás*. Plataforma Editorial, Colombia, 1er Edición.

[Nerio et al., 2018] Nerio, M., Pérez, J., and Wladimir, R. (2018). Reconocimiento de estados emocionales de personas mediante la voz utilizando algoritmos de aprendizaje de máquina. *Revista Venezolana de Computación*, 5(2):41–52.

[Nixon and Aguado, 2012] Nixon, M. S. and Aguado, A. S. (2012). *Feature Extraction Image Processing for Computer Vision*. Elsevier, Kidlington, Oxford, UK, 3th edition.

[Nolasco Martínez et al., 2013] Nolasco Martínez, J. J., Figueroa Hernández, F. C., and Vera Gutiérrez, F. (2013). Entrenamiento de una red neuronal en LabVIEW para la identificación en línea de un sistema dinámico. *Pistas Educativas*, 103:131–149.

[Ozdemir et al., 2019] Ozdemir, M. A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., and Akan, A. (2019). Real time emotion recognition from facial expressions

using cnn architecture. In *2019 Medical Technologies Congress (TIPTEKNO)*, pages 1–4, Izmir, Turkey.

[Pajares Martinsanz et al., 2010] Pajares Martinsanz, G., De La Cruz García, J. M., Ribeiro Seijas, , De Andrés y Toro, B., Martín Gómez, D., Oyen, D., Besada Portas, E., Rivas Rodríguez, J., Vega Sánchez, J. A., Conesa Muñoz, J. M., López Orosco, J. A., Martín Hernández, J. A., García-Alegre Sánchez, M. C., Guijarro Mata-García, M., Santos Peñas, M., Herrera Caro, P. J., Plis, S. M., Dormido Canto, S., and Burgos-Artizzu, X. P. (2010). *Aprendizaje Automático Un Enfoque Práctico*. RA-MA Editorial, Madrid, España. 1er Edición.

[Palmero et al., 2002] Palmero, F., Fernández-Abascal, E., Martínez Sánchez, F., and Choliz, M. (2002). *Psicología de la Motivación y la Emoción*. McGraw-Hill, Madrid, 1er edición.

[Peng et al., 2021] Peng, Z., Dang, J., Unoki, M., and Akagi, M. (2021). Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech. *Neural Networks*, 140:261–273.

[Perez Rosas et al., 2013] Perez Rosas, V., Mihalcea, R., and Morency, L. P. (2013). Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45.

[Pizer et al., 1987] Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368.

[Plutchik, 1962] Plutchik, R. (1962). *The emotions: Facts, theories and a new model*. Random House, New York, 1st edition.

[Pérez Espinosa and Reyes García, 2010] Pérez Espinosa, H. and Reyes García, C. (2010). Reconocimiento de emociones a partir de voz basado en un modelo emocional continuo. Technical report, Coordinación de Ciencias Computacionales INAOE.

[Rahdari et al., 2019] Rahdari, F., Rashedi, E., and Eftekhari, M. (2019). A Multimodal Emotion Recognition System Using Facial Landmark Analysis. *Iranian Journal of Science and Technology - Transactions of Electrical Engineering*, 43:171–189.

[Ramírez Cornejo and Pedrini, 2019] Ramírez Cornejo, J. Y. and Pedrini, H. (2019).

Bimodal emotion recognition based on audio and facial parts using deep convolutional neural networks. In *IEEE International Conference on Machine Learning and Applications*, pages 111–117, Boca Raton, FL, USA, USA.

[Rani and Yadav, 2018] Rani, P. and Yadav, M. B. (2018). A Survey on Gender and Emotion Recognition Using Voice. *International Journal of Recent Research Aspect*, 5(4):14–17.

[Reddy et al., 2019] Reddy, C. V. R., Reddy, U. S., and Kishore, K. V. K. (2019). Facial emotion recognition using nlpca and svm. *Traitement du Signal*, 36(1):13–22.

[Reddy and Kuchibhotla, 2019] Reddy, L. L. and Kuchibhotla, S. (2019). Survey on stress emotion recognition in speech. In *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 1–4, Greater Noida, India.

[Revina and Emmanuel, 2021] Revina, I. M. and Emmanuel, W. (2021). A survey on human face expression recognition techniques. *Journal of King Saud University - Computer and Information Sciences*, 33(6):619–628.

[Reza, 2004] Reza, A. M. (2004). Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38(1):35–44.

[Rodrıguez, 2017] Rodrıguez, L. G. M. (2017). *Mejora del contraste de imágenes a color utilizando un framework de optimización multiobjetivo*. PhD thesis, Universidad Nacional de Asunción.

[Roger Jang, 1997] Roger Jang, J.-S. (1997). *Neuro-Fuzzy and Soft Computing A Computional Approach to Learning and Machine Intelligence*. Prentice Hall, Upper Saddle River, New Jersey, 1st edition.

[Rolls, 1999] Rolls, E. (1999). The brain and emotions. *Trends in Cognitive Sciences*, 3(7):1–2.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65( ):386–480.

[Sahu et al., 2019] Sahu, S., Singh, A. K., Ghrera, S., Elhoseny, M., et al. (2019). An approach for de-noising and contrast enhancement of retinal fundus image using

clahe. *Optics and Laser Technology*, 110:87–98.

[Sajjad et al., 2020] Sajjad, M., Zahir, S., Ullah, A., Akhtar, Z., and Muhammad, K. (2020). Human behavior understanding in big multimedia data using cnn based facial expression recognition. *Mobile networks and applications*, 25(4):1611–1621.

[Sasaki, 1993] Sasaki, M. (1993). Facial expression and emotion. *Journal of Tokyo Medical University*, 76(2):219–223.

[Semwal et al., 2017] Semwal, N., Kumar, A., and Narayanan, S. (2017). Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pages 1–6, New Delhi, India.

[Shin et al., 2018] Shin, J., Maeng, J., and Kim, D.-H. (2018). Inner emotion recognition using multi bio-signals. In *2018 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, pages 206–212, JeJu, Korea (South).

[Shorten and Khoshgoftaar, 2019] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

[Sim et al., 2002] Sim, T., Baker, S., and Bsat, M. (2002). The CMU pose, illumination, and expression (pie) database. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 53–58, Washington, DC, USA.

[Singh and Singh, 2011] Singh, G. and Singh, B. (2011). Feature based method for human facial emotion detection using optical flow based analysis. *An International Journal of Engineering Sciences*, 4(1):363–372.

[Szeliski, 2010] Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

[Tan et al., 2020] Tan, H., Wu, G., Zhao, P., and Chen, Y. (2020). Spectrogram Analysis Via Self-Attention For Realizing Cross-Model Visual-Audio Generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4392–4396, Barcelona, España (Virtual).

[Tianmei et al., 2017] Tianmei, G., Jiwen, D., Henjian, L., and Yunxing, G. (2017). *Simple Convolutional Neural Network on Image Classification*. In *IEEE 2nd International Conference on Big Data Analysis*, pages 721–724, Beijing, China.

[Tortosa Gil and Mayor Martínez, 1992] Tortosa Gil, F. and Mayor Martínez, L. (1992). Watson y la psicología de las emociones: Evolución de una idea. *Psicothema*, 4(1):297–315.

[Trigeorgis et al., 2017] Trigeorgis, G., Nicolaou, M. A., and Schuller, W. (2017). End to End Multimodal Emotion Recognition. *Ieee Journal of Selected Topics in Signal Processing*, 11(8):1301–1309.

[Vasilev et al., 2019] Vasilev, I., Slater, D., Spacagna, G., Roelants, P., and Zocca, V. (2019). *Python Deep Learning: Exploring deep learning techniques and neural network architectures with Pytorch, Keras, and TensorFlow*. Packt Publishing Ltd.

[Viola and Jones, 2004] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

[Wagner et al., 2011] Wagner, J., Lingenfelser, F., André, E., and Kim, J. (2011). Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4):206–218.

[Wang et al., 2017] Wang, J., Perez, L., et al. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vision Recognition*, 11:1–8.

[Wang et al., 2020] Wang, X., Chen, X., and Cao, C. (2020). Human emotion recognition by optimally fusing facial expression and speech feature. *Signal Processing: Image Communication*, 84:115831–115837.

[Wang and Guan, 2008] Wang, Y. and Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE transactions on multimedia*, 10(5):936–946.

[Wootaek et al., 2016] Wootaek, L., Daeyoung, J., and Lee, T. (2016). Speech Emotion Recognition using Convolutional Recurrent Neural Networks and Spectrograms. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4, London, ON, Canada.

[Wu et al., 2014] Wu, C.-H., Lin, J.-C., and Wei, W.-L. (2014). Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing*, 3(12):1–18.

[Xu et al., 2014] Xu, X., Li, Y., Xu, X., Wen, Z., Che, H., Liu, S., and Tao, J. (2014). Survey on discriminative feature selection for speech emotion recognition. In *The*

9th International Symposium on Chinese Spoken Language Processing, pages 345–349, Singapore.

[Yadav et al., 2014] Yadav, G., Maheshwari, S., and Agarwal, A. (2014). Contrast limited adaptive histogram equalization based enhancement for real time video system. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2392–2397, Delhi, India.

[Yang and Ma, 2019] Yang, J. and Ma, J. (2019). Feed-forward neural network training using sparse representation. *Expert Systems with Applications*, 116:255–264.

[Zhang et al., 2018] Zhang, S., Zhang, S., Huang, T., Gao, W., and Tian, Q. (2018). Learning Affective Features with a Hybrid Deep Model for Audio-Visual Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):1–14.

[Zhao et al., 2019] Zhao, J., Mao, X., and Chen, L. (2019). Speech emotion recognition using deep 1d and 2d CNN LSTM networks. *Biomedical Signal Processing and Control*, 47:312–323.

[Zhao et al., 2018] Zhao, J., Mao, X., and Zhang, J. (2018). Learning deep facial expression features from image and optical flow sequences using 3d cnn. *The Visual Computer*, 34(10):1461–1475.

Anexos

A. Manual de Usuario

En este anexo se indica cómo usar la biblioteca desarrollada para realizar la ejecución de los modelos obtenidos para el reconocimiento de emociones con la base de datos UTeMo, los cuales se implementaron en el lenguaje de programación Python. La estructura general del directorio raíz del proyecto se muestra en la Figura A.1.

La descripción del contenido del directorio raíz se presenta a continuación.

- En la carpeta `Bases_de_datos` se incluyen los datos de prueba utilizados en las dos modalidades de información (voz e imágenes).

- En la carpeta `Modelos` se incluyen los modelos de voz, expresiones faciales e híbrido. Cabe mencionar que sólo se incluyeron los modelos con mejor desempeño, es decir, los modelos implementados con cocleogramas (para voz) y

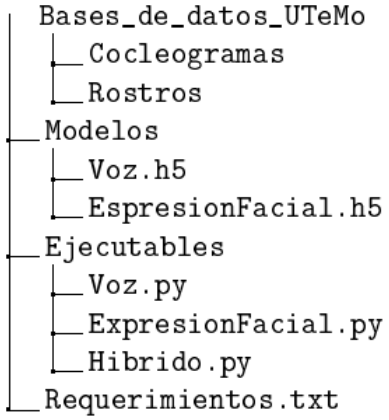


Figura A.1: Estructura del directorio raíz del proyecto

expresiones faciales.

- En la carpeta `Ejecutables` se encuentran los scripts de Python para reconocimiento de emociones. Se incluyeron tres scripts: uno para las predicciones de voz, otro para las predicciones con expresiones faciales, y finalmente un script para la predicción híbrida.

- Finalmente, el archivo `Requerimientos.txt` contiene el listado de dependencias del proyecto.

A.1 Instalación de dependencias

Para la instalación de las dependencias requeridas y asegurar el correcto funcionamiento de la biblioteca desarrollada, se hace uso del Instalador de Paquetes para Python (PIP, por sus siglas en inglés), el cual facilita la instalación de bibliotecas. Para iniciar el proceso de instalación, basta con ejecutar los siguientes comandos:

```
$ sudo apt install python3-pip
```

```
$ sudo pip3 install -r Requerimientos.txt
```

El primer comando instala la herramienta PIP para ser usada con Python3; mientras que el segundo instala todas las bibliotecas que son dependencias del proyecto. Como referencia, en la Figura A.2 se muestra la descarga de las dependencias

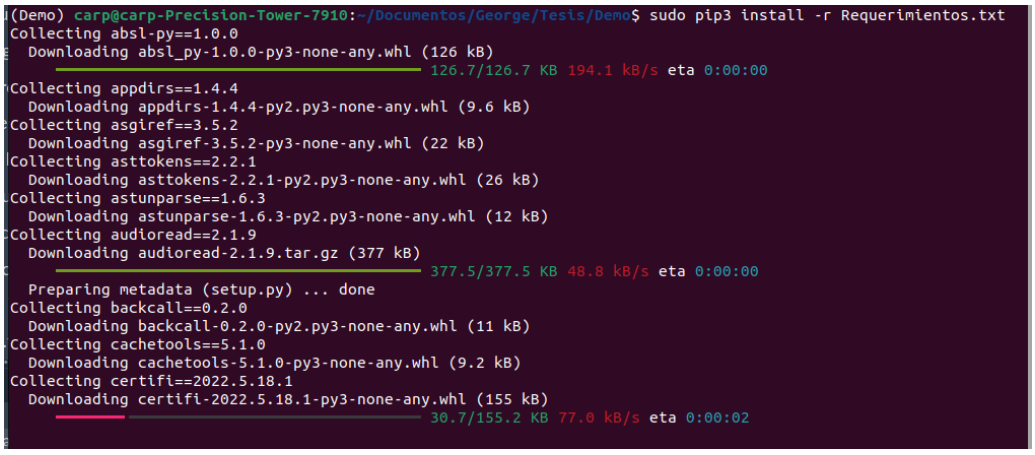


Figura A.2: Ejemplo de la correcta ejecución de los comandos para la instalación de dependencias

una vez ejecutados ambos comandos.

A.2 Ejemplo práctico

Para iniciar la ejecución, basta con ejecutar el script de python deseado con el siguiente comando

```
$ python3 Ejecutables/{script}
```

donde `{script}` puede tomar los valores de `Voz.py`, `ExpresionFacial.py` o `Hibrido.py`. Por ejemplo, si se desea ejecutar el modelo de voz, se debe ejecutar el comando:

```
$ python3 Ejecutables/Voz.py
```

El comando anterior devolverá la exactitud de clasificación, matriz de confusión y otras medidas de rendimiento, como se muestra en la Figura A.3. De manera similar, en las Figuras A.4 y A.5 se muestran los resultados al ejecutar los comandos `python3 Ejecutables/ExpresionFacial.py` y `python3 Ejecutables/Hibrido.py` respectivamente.

```
Found 183 images belonging to 7 classes.
Modelo cargado
6/6 [==============================] - 2s 37ms/step
6/6 [==============================] - 0s 31ms/step - loss: 0.2011 - accuracy: 0.9508

Resultado en las pruebas:  0.9508196711540222


['Asco', 'Felicidad', 'Ira', 'Miedo', 'Neutral', 'Sorpresa', 'Tristeza']
[[24  0  0  0  0  0  1]
 [ 0 24  0  0  0  0  1]
 [ 0  0 25  0  0  0  0]
 [ 0  0  0 23  0  1  1]
 [ 0  0  0  0 26  0  0]
 [ 0  2  0  1  0 24  1]
 [ 0  0  0  0  0  1 28]]


              precision    recall  f1-score   support

        Asco       1.00      0.96      0.98        25
   Felicidad       0.92      0.96      0.94        25
         Ira       1.00      1.00      1.00        25
      Miedo        0.96      0.92      0.94        25
     Neutral       1.00      1.00      1.00        26
    Sorpresa       0.92      0.86      0.89        28
    Tristeza       0.88      0.97      0.92        29

    accuracy                           0.95       183
   macro avg       0.95      0.95      0.95       183
weighted avg       0.95      0.95      0.95       183
```

Figura A.3: Ejemplo de la ejecución del archivo Voz.py

```
Found 817 images belonging to 7 classes.
Modelo cargado
26/26 [==============================] - 5s 95ms/step
26/26 [==============================] - 3s 92ms/step - loss: 0.1742 - accuracy: 0.9596

Resultado en las pruebas:  0.9596083164215088


['Asco', 'Felicidad', 'Ira', 'Miedo', 'Neutral', 'Sorpresa', 'Tristeza']
[[114   0   0   6   0   0   0]
 [  1 101   0   0   0   4   0]
 [  0   0  96   1   0   8   0]
 [  1   0   1  99   0   0   1]
 [  0   0   0   0 130   0   0]
 [  1   3   2   0   0 120   1]
 [  2   0   1   0   0   0 124]]


              precision    recall  f1-score   support

       Asco       0.96      0.95      0.95       120
   Felicidad       0.97      0.95      0.96       106
         Ira       0.96      0.91      0.94       105
      Miedo       0.93      0.97      0.95       102
    Neutral       1.00      1.00      1.00       130
   Sorpresa       0.91      0.94      0.93       127
   Tristeza       0.98      0.98      0.98       127

   accuracy                           0.96       817
  macro avg       0.96      0.96      0.96       817
weighted avg       0.96      0.96      0.96       817
```

Figura A.4: Ejemplo de la ejecución del archivo ExpresionFacial.py

```
Found 183 images belonging to 7 classes.
6/6 [==============================] - 2s 38ms/step
Found 817 images belonging to 7 classes.
26/26 [==============================] - 4s 104ms/step

Exactitud de clasificación:  0.99265605875153


['Asco', 'Felicidad', 'Ira', 'Miedo', 'Neutral', 'Sorpresa', 'Tristeza']
[[119   0   0   1   0   0   0]
 [  0 105   0   0   0   1   0]
 [  0   0 105   0   0   0   0]
 [  0   0   0 102   0   0   0]
 [  0   0   0   0 130   0   0]
 [  0   1   1   1   0 123   1]
 [  0   0   0   0   0   0 127]]


              precision    recall  f1-score   support

       Asco       1.00      0.99      1.00       120
   Felicidad       0.99      0.99      0.99       106
         Ira       0.99      1.00      1.00       105
      Miedo       0.98      1.00      0.99       102
    Neutral       1.00      1.00      1.00       130
   Sorpresa       0.99      0.97      0.98       127
   Tristeza       0.99      1.00      1.00       127

   accuracy                           0.99       817
  macro avg       0.99      0.99      0.99       817
weighted avg       0.99      0.99      0.99       817
```

Figura A.5: Ejemplo de la ejecución del archivo Hibrido.py