UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA
DIVISIÓN DE ESTUDIOS DE POSGRADO

Tesis
Para obtener el grado de:
MAESTRA EN ROBÓTICA

SISTEMA MULTIMODAL DE RECONOCIMIENTO DE EMOCIONES EMPLEANDO LA INTERACCIÓN CON EL ROBOT HUMANOIDE NAO.

Presenta:

Ing. Vanessa Cristell Torres López

Director de tesis:

Dr. José Aníbal Arias Aguilar

Co-Director de tesis:

Dr. Eduardo Sánchez Soto

Heroica Ciudad de Huajuapan de León, Oaxaca, agosto 2024

DEDICATORIA

Esta tesis está dedicada con todo mi amor y gratitud a las personas que han sido mi fortaleza y mi inspiración a lo largo de este arduo pero gratificante camino.

A mi esposo:

Por su amor incondicional, su apoyo constante y su paciencia infinita. Gracias por estar a mi lado en cada paso de este viaje, por creer en mí y por animarme a seguir adelante incluso en los momentos más difíciles.

A mis hijos:

Ustedes son mi mayor motivación. Sus sonrisas y abrazos me dieron la fuerza para perseverar y alcanzar mis metas. Esta tesis es también para ustedes, como un recordatorio de que, con esfuerzo y dedicación, todo es posible.

A mi mamá:

Por su amor incondicional, sus sabios consejos y su ejemplo de tenacidad y dedicación. Mamá, tu apoyo ha sido inquebrantable y siempre me has mostrado el camino con tu fuerza y sabiduría.

A mi abuelito:

Por ser un modelo de perseverancia y esfuerzo. Abuelito, tus historias y enseñanzas me han inspirado a seguir adelante y a nunca rendirme.

A Dios:

Por darme la fuerza, la sabiduría y las oportunidades para llegar hasta aquí. Tu guía y bendiciones han sido esenciales en cada paso de este camino.

Con todo mi amor y agradecimiento,

Vanessa Cristell Torres López

AGRADECIMIENTOS

Al concluir este importante proyecto de tesis de maestría, deseo expresar mi más profundo agradecimiento a todas las personas que hicieron posible este trabajo.

A mis directores de tesis:

Dr. José Aníbal Arias Aguilar y Dr. Eduardo Sánchez Soto por su invaluable orientación, apoyo constante y por compartir su vasto conocimiento y experiencia. Su paciencia y sus consejos fueron esenciales para el desarrollo y finalización de esta investigación.

A mis sinodales:

Dr. Ignacio Arroyo Fernández, Dr. Oscar Ramírez y Dr. Rosebet Miranda Luna, Dr. Miguel Alberto Domínguez Gurría por su tiempo, dedicación y aportes críticos que enriquecieron significativamente este trabajo. Sus sugerencias y comentarios fueron fundamentales para alcanzar el nivel académico deseado.

Al club de teatro:

Por su cooperación en la creación de la base de datos emocional con el robot NAO. Su disposición y colaboración fueron clave para la recopilación de datos esenciales para esta investigación.

Al equipo de multimedios:

Por su excelente trabajo en las grabaciones de las sesiones con NAO. Su profesionalismo y habilidades técnicas aseguraron que todos los detalles fueran capturados con precisión, lo cual fue crucial para el análisis posterior.

A todos ustedes, les estoy profundamente agradecido por su compromiso y apoyo, que fueron indispensables para la realización de esta tesis. Su colaboración y contribuciones han dejado una huella imborrable en mi formación académica y personal.

Con gratitud,

Vanessa Cristell Torres López.

RESUMEN

Actualmente, el reconocimiento de emociones, es un tema de investigación que se encuentra en auge debido a la necesidad de establecer una Interacción Hombre-Robot (IHR) natural. Esta investigación, describe el desarrollo de dos sistemas unimodales y un sistema bimodal para el reconocimiento de emociones bajo dos señales de entrada: imagen y voz.

En la primera fase de esta investigación, se implementaron diversos modelos de Redes Neuronales Convolucionales (CNN, por sus siglas en inglés) con cada una de las señales de entrada, empleando bases de datos preexistentes en las fases de entrenamiento, validación y prueba con el fin de ajustar el modelo para mejorar su desempeño. En la segunda fase se ponderaron los resultados obtenidos previamente, obteniendo un sistema bimodal que contempla las dos señales de entrada. Para la última fase las CNN se vuelven a entrenar y a ajustar, en esta ocasión, con una base de datos propia obtenida a partir de la interacción con el robot humanoide NAO.

El objetivo de generar una base de datos a partir de la interacción con NAO consiste en obtener datos más naturales al reducir la sobreactuación de las emociones estando frente a un robot de aspecto humanoide.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Capítulo 1

INTRODUCCIÓN

Los inicios de la robótica se remontan al primer tercio del siglo XX, cuando se desarrolló en conjunto con las diferentes ramas de la ingeniería, las cuales brindaron el impuso necesario para el diseño, manufactura y control de máquinas que efectúan tareas definidas por el usuario (Sánchez Martín et al., 2007).

Es así, que una manera de clasificar a los robots es de acuerdo al medio en el que interactúan y la actividad que realizan. De este modo, encontramos seis tipos de robots (Siciliano & Khatib, 2016):

- Robots industriales
- Robots militares
- Robots de entretenimiento
- Robots en la industria médica
- Robots de servicio
- Robots sociales

En esta investigación, nos enfocaremos en los robots sociales. Un robot social, es aquel que interactúa y se comunica con las personas de una manera sencilla y agradable, siguiendo comportamientos, patrones y normas sociales. La complejidad de un robot social está dada por la aceptación de estos con la sociedad, puesto que se requiere sean capaces de comunicarse con las personas manteniendo un lenguaje de alto nivel, por lo que deben interpretar el habla humana, además, deben reconocer las expresiones faciales, los gestos y las acciones humanas, finalmente, deben interpretar la conducta social de las personas a través de la construcción de elaborados modelos cognitivos-afectivos (Breazeal, 2004).

1.1 Emociones

Una emoción es definida como el proceso que se activa cuando el organismo detecta algún cambio en su entorno con el fin de poner en marcha los recursos a su alcance para controlar la situación. Por ejemplo, el miedo provoca un aumento del latido cardiaco que hace que llegue más sangre a los músculos favoreciendo la respuesta de huida (Ekman, et.al, 1971) (Asociación Española contra el Cáncer, 2010).

Las emociones se clasifican en básicas y complejas, donde las últimas son resultado de la coexistencia de dos o más emociones básicas. Se conocen seis emociones básicas, que se describen a continuación (Asociación Española contra el Cáncer, 2010, p. 22) (Chóliz Montañés, 2005, p. 45):

- Miedo: Anticipación de una amenaza o peligro (real o imaginario) que produce ansiedad, incertidumbre e inseguridad.
- Sorpresa: Sobresalto, asombro, desconcierto. Es una emoción transitoria.
- Aversión: Disgusto hacia aquello que tenemos en frente.
- Ira: Rabia, enojo que aparece cuando las cosas no salen como queremos.
- Tristeza: Pena, soledad, pesimismo ante la pérdida de algo importante.
- Alegría: Sensación de bienestar y de seguridad.

1.2 Antecedes

El reconocimiento de emociones implica la identificación y clasificación de estados emocionales a partir de datos sensoriales. Las emociones se pueden manifestar de diversas formas, incluyendo expresiones faciales, tono de voz, gestos corporales y fisiología. Tradicionalmente, este reconocimiento se ha llevado a cabo mediante técnicas de procesamiento de señales y aprendizaje automático que, aunque efectivas, enfrentan limitaciones en términos de escalabilidad y precisión. El reconocimiento de emociones se efectúa a través de diferentes medios, tales como la voz, y el reconocimiento de expresiones faciales.

1.2.1 Importancia del Reconocimiento de Emociones

El reconocimiento de emociones tiene aplicaciones en múltiples campos:

- Educación: Facilita la creación de sistemas educativos adaptativos.
- Interacción Hombre-Robot: Mejora la comunicación entre humanos y dispositivos inteligentes.
- Marketing: Permite analizar las reacciones emocionales de los consumidores.
- Salud Mental: Ayuda en el diagnóstico y seguimiento de trastornos emocionales.

1.2.2 Métodos Tradicionales de Reconocimientos de Emociones vs. Métodos Basados en CNN

Los métodos tradicionales para el reconocimiento de emociones a menudo implican el uso de características manuales y clasificadores basados en reglas o aprendizaje automático. Estos métodos, aunque efectivos en ciertos contextos, son limitados en su capacidad para manejar variaciones complejas y sutilezas en las expresiones emocionales.

Las CNN, por otro lado, aprenden características directamente a partir de los datos sin necesidad de ingeniería manual. Esto no solo mejora la precisión, sino que también permite una mejor escalabilidad y adaptación a nuevos datos. Las CNN pueden capturar patrones jerárquicos y complejos en los datos de entrada, lo que las hace especialmente adecuadas para el reconocimiento de emociones en imágenes y señales de audio (LeCun, et.al, 2015).

A continuación, se muestra una reseña de dichos trabajos.

Reconocimiento de Emociones en Expresiones Faciales

Las investigaciones mostradas a continuación muestran el resultado del uso de CNN como principal clasificador de emociones detectadas en expresiones faciales. Se recopila la información acerca de la arquitectura final empleada, la base de datos utilizada y el porcentaje de precisión obtenido en cada investigación.

Yu, Z., y Zhang, C. (2015) en su participación en el *FER Challenge* plantean la arquitectura de CNN mostrada en la Figura 1 para clasificar siete emociones: felicidad, tristeza, miedo, sorpresa, disgusto, ira y neutralidad en el conjunto de datos de imágenes estáticas SFEW obteniendo una precisión de 61% en el conjunto de imágenes de prueba.

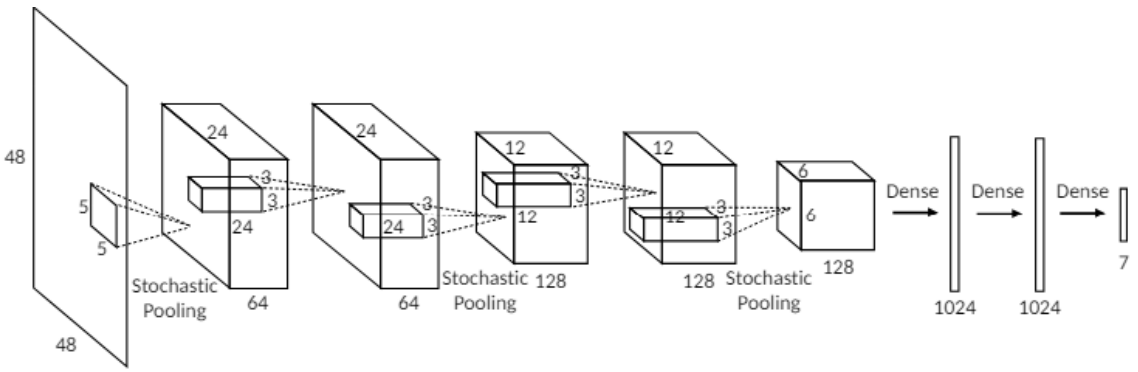


Figura 1. Arquitectura CNN simple para reconocimiento de expresiones faciales. Tomado de Yu.Z. y Zhang, C. (2015)

Rajan et al. (2020), clasifican siete emociones: felicidad, tristeza, miedo, sorpresa, disgusto, ira y neutralidad. En su investigación, realizan dos preprocesamientos diferentes a la imagen de la expresión facial a fin de preservar información sutil sobre los bordes de cada imagen. Cada uno de los dos grupos ingresan a una arquitectura de CNN individual que se fusionan e integran con una capa de *Long Short-Term Memory* (LSTM) que extrae las relaciones temporales de los fotogramas sucesivos. La arquitectura propuesta, mostrada en la Figura 2 fue probada en las bases de datos MMI, SFEW y una base de datos propia obteniendo precisiones de 80%, 43% y 95% respectivamente.

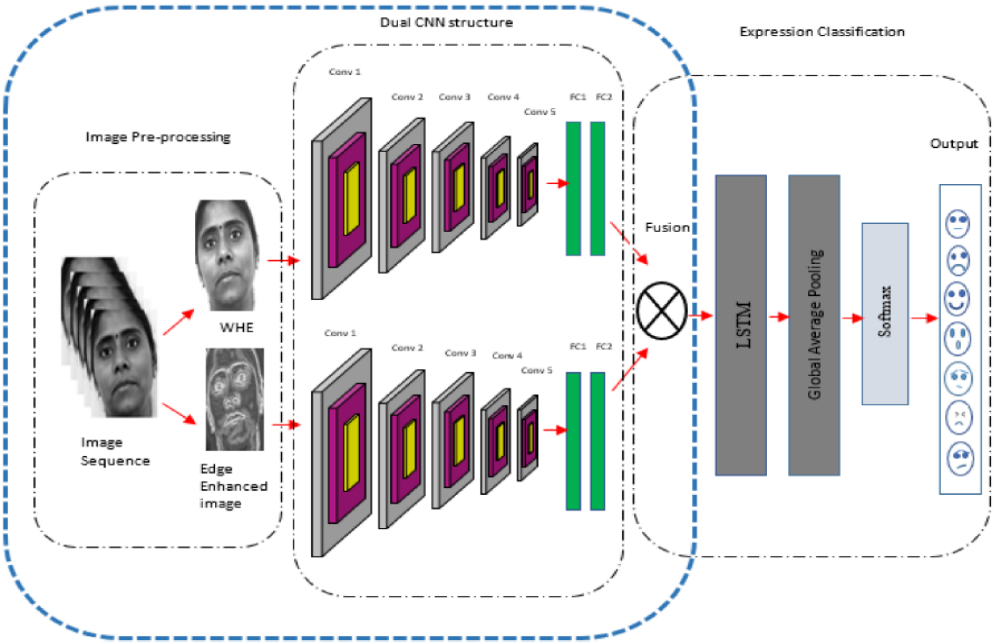


Figura 2. Arquitectura CNN + LSTM. Tomada de Rajan et.al (2020)

Akram, Alhajlah y Mahmood (2023), evalúan diferentes algoritmos de *Machine Leaning* (ML), tales como: *Support Vector Machine* (SVM), *Desicion Tree, Linear Discriminant Analysis* (LDA), entre otros, a la salida de las arquitecturas de AlexNet , como se muestra en la Figura 3, y VGG-16 para la clasificación de tres emociones: positiva, negativa y neutral en la base de datos MLF-W-FER logrando mejorar la precisión inicial de las arquitecturas en 7%-9% con lo que se obtienen desempeños promedio de 65%.

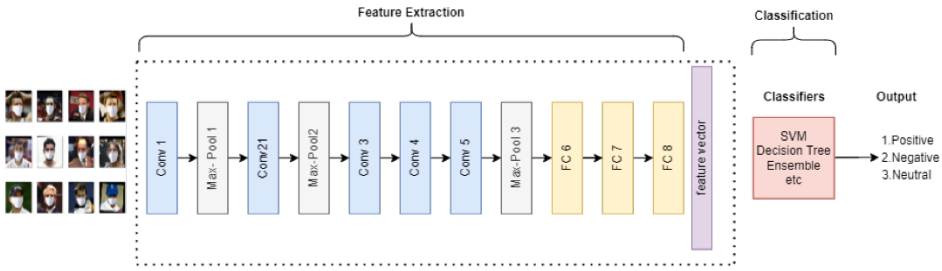


Figura 3. Evaluación de técnicas de ML con la arquitectura AlexNet. Tomada de Akram, Alhajlah y Mahmood (2023)

Obaid y Alrammahi (2023), presentan un sistema de reconocimiento de expresiones faciales que clasifica siete emociones: felicidad, tristeza, miedo, sorpresa, disgusto, ira y neutralidad utilizando una CNN híbrida mostrada en la Figura 4. Este modelo combina una red CNN para procesar imágenes faciales estáticas y una red de creencias profundas (DBN) para integrar características espaciales y temporales. El modelo alcanzó un alto rendimiento de reconocimiento en las bases de datos JaFFE, KDEF y RaFD, con 98%, 95% y 98% respectivamente.

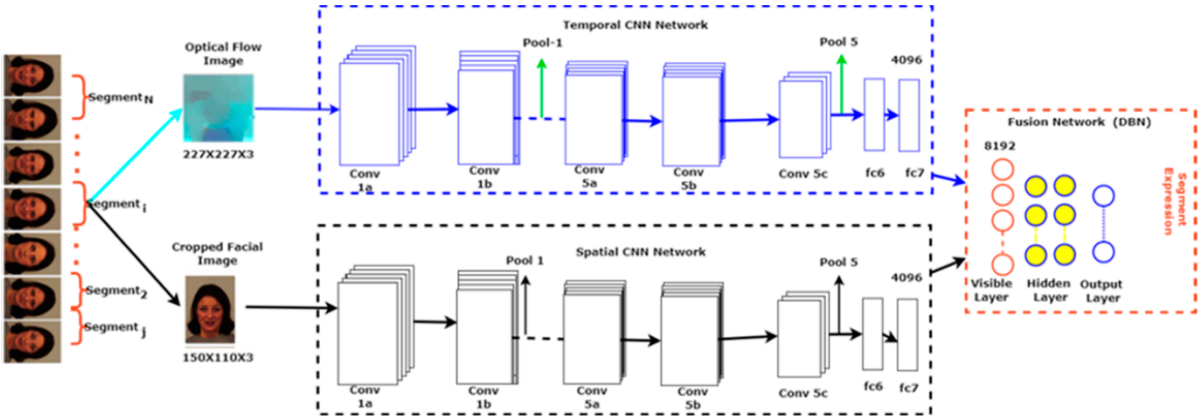


Figura 4. Arquitectura CNN híbrida para reconocimiento de expresiones faciales. Toma de Obaid y Alrammahi (2023)

A partir de las investigaciones citadas, se observa que la precisión en los modelos de CNN para la clasificación de expresiones faciales mejora considerablemente cuando son fusionados con otras técnicas de ML y *Deep Learning* (DL)

Reconocimiento de Emociones en Voz

Murugan (2020) propone el modelo de CNN mostrado en la Figura 5, el modelo fue entrenado y evaluado con la base de datos RAVDESS con ocho emociones: felicidad, tristeza, enojo, miedo, disgusto, sorpresa, desprecio y neutralidad. En su investigación, transforma los audios de la base de datos en espectrogramas y gráficas de formas de onda

empleando la librería librosa de Python mismos que representan la entrada de la CNN. El modelo propuesto obtuvo un rendimiento del 71%.

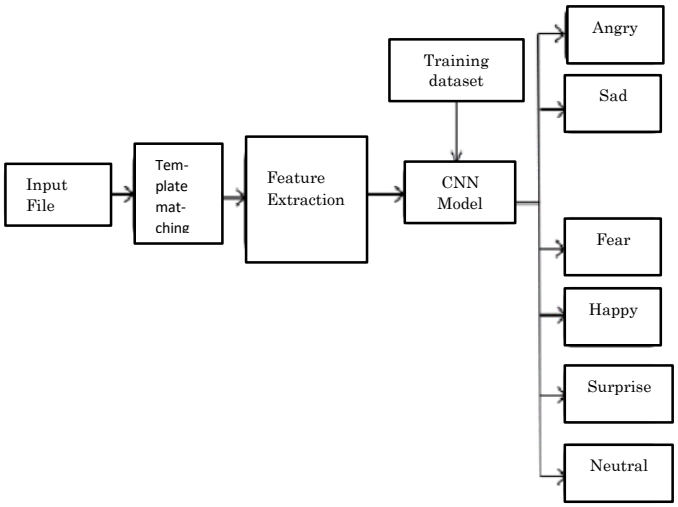


Figura 5. Arquitectura de CNN para el reconocimiento de emociones en voz. Tomado de Murugan (2020)

Zhao, Zou y Zhang (2023) exploraron el uso de CNN *y Multi-Head Convolutional Transformers*, tal como se muestra en la Figura 6, para el reconocimiento emociones en el habla a partir de los espectrogramas de los audios contenidos en las bases de datos RAVDESS y IEMOCAP (cinco emociones: felicidad, ira, tristeza, miedo y neutralidad). El modelo alcanzó un rendimiento del 82% y 79% respectivamente.

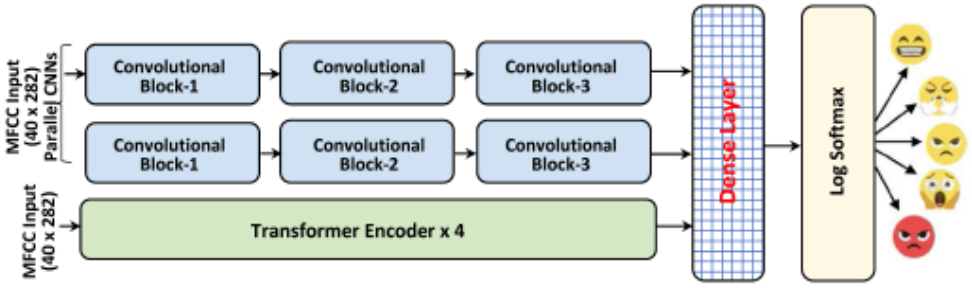


Figura 6. Arquitectura CNN y Multi-Head Convolutional Transformers. Tomada de Zhao, Zou y Zhang (2023)

Kim y Lee (2023) en su investigación proponen una nueva arquitectura que combina *Bidirectional Long Short-Term Memory* (BiLSTM), *Transformer* y CNN 2D, Figura 7. En su investigación, transforma las señales de audio en espectrogramas que son entradas adecuadas para la CNN 2D. El sistema propuesto en el artículo se entrenó y evaluó para clasificar siete emociones: felicidad, tristeza, enojo, miedo, sorpresa, asco y neutralidad utilizando las bases de datos Emo-DB y RAVDESS, obteniendo precisiones promedio de 89% y 70% respectivamente al aplicar validación cruzada sobre 10 *folds*.

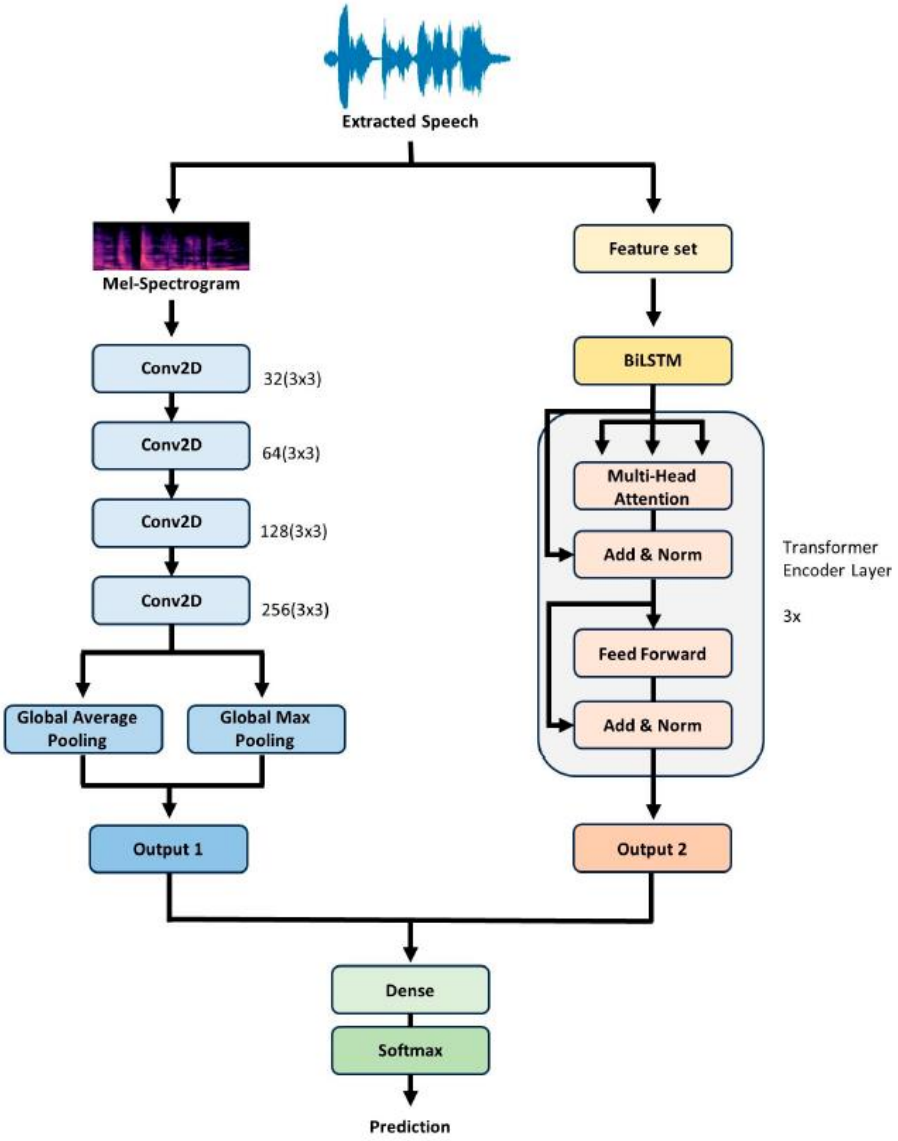


Figura 7. Arquitectura BiLSTM-Transformer-CNN2D para reconocimiento de emociones en habla. Tomada de Kim y Lee (2023)

Muhammad (2024) presenta el modelo CNN 3D, ilustrado en la Figura 8, con fusión de funciones múltiples al incorporar tres técnicas de extracción de características distintas, como los Coeficientes Cepstrales de Frecuencia Mel (MFCC), *Chroma Shift* y un espectrograma Mel. El modelo se probó y entrenó utilizando las bases de datos SUBESCO, con 8 emociones (aburrimiento, fatiga, confusión, frustración, alegría leve, tristeza leve, ansiedad leve y desprecio leve) y RAVDESS obteniendo precisiones de 96 y 89% respectivamente.

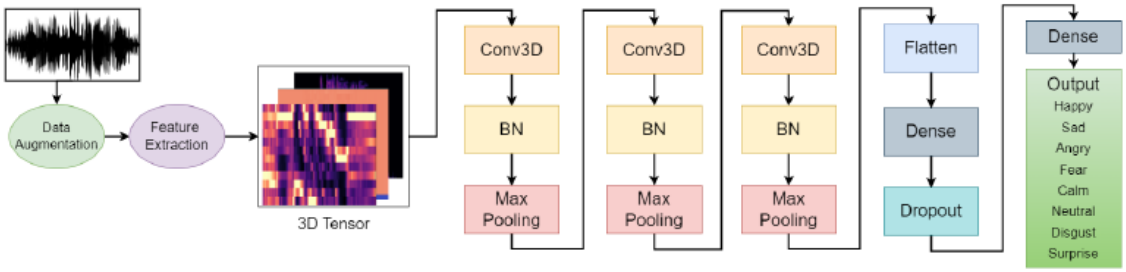


Figura 8. Arquitectura CNN 3D con fusión de funciones múltiples. Tomada de Muhammad (2024)

La revisión de antecedentes para el reconocimiento de emociones en el habla (SER, *por sus siglas en inglés*) muestra que la base de datos utilizada con mayor frecuencia es RAVDESS. Se observa además que se obtiene una mejor precisión al momento de clasificar emociones analizando señales de habla en comparación con imágenes de rostros humanos puesto que una CNN simple siendo entrenada con los espectrogramas muestra un porcentaje de precisión superior al obtenido de modelos de CNN simples que clasifican expresiones faciales.

Reconocimiento Multimodal de Emociones.

Pan (2023) en su investigación analiza el habla, expresiones faciales y electroencefalogramas (EEG) a través de un Reconocimiento Multimodal de Emociones (MER, *por sus siglas en inglés)* basado en el aprendizaje llamado *Deep-Emotion.* El modelo propuesto contiene una rama por cada señal de entrada empleando el modelo *GhostNet* para la rama de las expresiones faciales, un algoritmo L*ightweight Fully*

Convolutional Neural Network (LFCNN) para la señal del habla y un modelo LSTM en el caso de los EEG. La arquitectura completa se muestra en la Figura 9. Pan reporta una precisión promedio de 96% en el modelo propuesto al utilizar las bases de datos CK+, Emo-DB y MAHNOB-HCI.

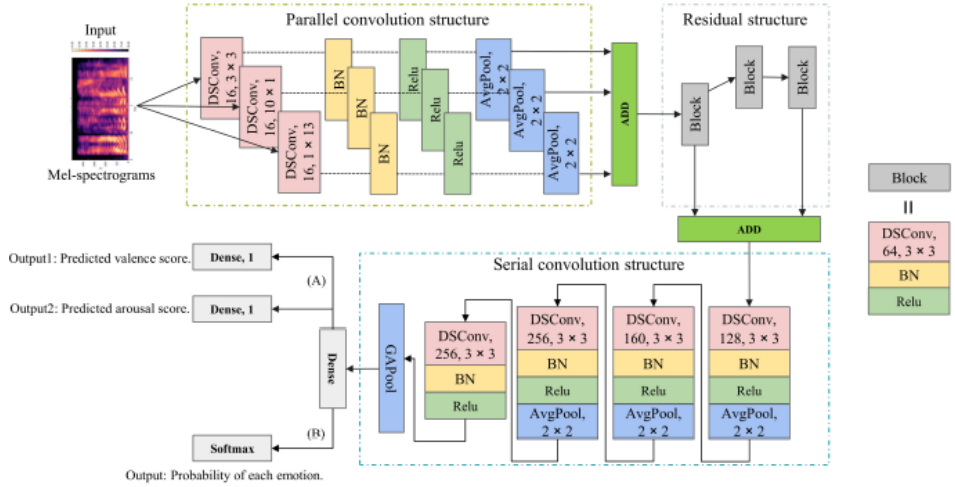


Figura 9. Arquitectura LFCC-LSTM para el reconocimiento de emociones en voz. Tomada de Pan (2023)

Enfoque de la investigación actual

Así como las arquitecturas que emplean más de un algoritmo de *Deep-Learning* para la clasificación de emociones en un tipo de señal de entrada muestran desempeños superiores a aquellos modelos que contemplan únicamente CNN, MER supone una opción de mejora en la precisión para la clasificación de emociones, es por ello que se propone el desarrollo de tres sistemas de reconocimiento de emociones basados en CNN que consistan en el análisis de dos señales: voz e imagen, a fin de evaluar su desempeño y determinar si un mayor número de señales incrementa la precisión del sistema.

Se implementarán dos sistemas unimodales que analicen cada una de las señales de manera independiente y un sistema bimodal que integre voz e imagen.

La fase de entrenamiento de los modelos resultantes se efectuará con dos tipos de bases de datos diferentes: primero, con la integración de los *datasets* existentes CK+ y FER 2013 para el análisis de expresiones faciales y RAVDESS en el caso del estudio del habla; por último, los modelos serán entrenados y validados con una base de datos propia

de dos ramas creada a partir de la interacción con el robot humanoide NAO a fin de evaluar el impacto que supone conversar con un robot antropomórfico.

1.3 Planteamiento del Problema

En la actualidad, el reconocimiento de emociones en los robots sociales representa uno de los retos más importantes para la inteligencia artificial dado que se pretende que además de ejecutar la tarea que se le fue asignada, el robot pueda establecer una relación más afectiva con el humano de acuerdo a su estado de ánimo. Sin embargo, el problema va más allá del simple hecho de establecer algoritmos de reconocimiento de emociones, pues se sabe que la emoción del ser humano puede verse afectada negativamente cuando existe rechazo hacia el robot derivado de la forma poco antropomórfica del ser, de acuerdo al efecto *Uncanny Valley* (Mori, et.al, 2012)

Por lo mencionado anteriormente, el desarrollo de un sistema MER empleando la interacción con el robot de aspecto humanoide NAO proporcionará una IHR de una forma más natural, puesto que el robot interpretará la emoción del usuario para poder establecer una comunicación afectiva con el mismo.

1.4 Justificación

Dentro de la robótica, uno de los aspectos que ha cobrado interés es la comunicación con el ser humano. Actualmente se busca que los robots sean capaces de entablar conversaciones afectivas de alto nivel con el usuario tomando en cuenta las emociones del mismo. Es entonces donde se vuelve oportuno y conveniente el diseño de un sistema MER que involucre señales de audio e imagen con el fin de comparar los resultados del mismo con algoritmos de una señal para determinar si la tasa de reconocimiento se mejora con el uso del mismo.

La investigación propuesta, cobra relevancia en la robótica social pues ampliaría el panorama de la IHR ya que además de reconocer las emociones en el usuario, los robots deben ser capaces de crear un vínculo afectivo.

1.5 Hipótesis

"Un sistema de reconocimiento multimodal de emociones que integre las señales de voz e imagen, bajo la interacción con el robot humanoide NAO incrementará el desempeño que el sistema obtiene al tratar las señales de manera unimodal".

1.6 Objetivos

1.6.1 Objetivo General

Diseñar un sistema MER que integre señales de audio e imagen, empleando al robot humanoide NAO como medio de adquisición de datos para el mejoramiento de la tasa de reconocimiento de emociones en los robots sociales.

1.6.2 Objetivos específicos

- Establecer una comunicación natural entre el robot NAO y el usuario a fin de eliminar errores derivados del rechazo hacia el robot.
- Crear una base de datos con los sensores de voz e imagen propios del robot humanoide NAO.
- Realizar dos sistemas unimodales de reconocimiento de emociones que evalúen de manera independiente las señales de voz e imagen.
- Implementar un sistema bimodal de reconocimiento de emociones evaluando las señales de voz e imagen.
- Comparar el comportamiento de los sistemas desarrollados a fin de evaluar si el número de señales y la interacción con el robot NAO influye en el desempeño del sistema de reconocimiento de emociones.

1.7 Limitaciones de la Tesis

- Debido a su baja capacidad de cómputo, el robot humanoide NAO únicamente se utilizará como medio de adquisición de datos. El procesamiento de las señales se efectuará en un equipo de cómputo externo.
- Las pruebas se efectuarán en un ambiente controlado, con personas de entre 18 a 25 años.

Capítulo 2

MARCO TEÓRICO

A continuación, se presentan los fundamentos teóricos que servirán de base a lo largo del desarrollo de la investigación propuesta.

2.1 Redes Neuronales Convolucionales (CNN)

La capacidad de una máquina para reconocer y responder a las emociones humanas puede mejorar significativamente la experiencia del usuario y la efectividad de diversas aplicaciones (Poria et.al, 2017). Con los avances en inteligencia artificial y el aprendizaje profundo, las Redes Neuronales Convolucionales (CNN) han emergido como una herramienta poderosa para mejorar la precisión y eficiencia en el reconocimiento de emociones, especialmente a través del análisis de expresiones faciales y señales de voz (Han, et.al, 2014).

Las CNN son un tipo de red neuronal diseñada para procesar datos con una estructura bidimensional, como las imágenes y los espectrogramas de audio. La arquitectura básica de una CNN incluye (Goodfellow, et. al, 2016):

- Capas Convolucionales: Las capas convolucionales son el núcleo de las CNN. Utilizan filtros (kernels) que recorren la imagen de entrada para crear mapas de características. Cada filtro aprende a detectar diferentes características, como bordes, texturas o patrones complejos.
- Capas de Pooling: Estas capas reducen la dimensionalidad de los mapas de características, lo que ayuda a disminuir el tiempo de computación y a controlar el sobreajuste. El pooling puede ser de tipo máximo (*Max Pooling*), promedio *(Average Pooling)* entre otros.

- Capas Completamente Conectadas: Al final de la red, las capas completamente conectadas *(Fully Connected Layers)* actúan como un clasificador. Cada neurona en estas capas está conectada a todas las neuronas en la capa anterior, lo que permite la combinación de características para realizar la clasificación final.
- Funciones de Activación: Las funciones de activación se utilizan para introducir no linealidades en la red, permitiendo que la CNN aprenda relaciones complejas en los datos.
- Regularización: Técnicas como Dropout se utilizan para prevenir el sobreajuste durante el entrenamiento, desactivando aleatoriamente un porcentaje de neuronas en cada paso.

Estas capas permiten a las CNN capturar tanto las características espaciales (en el caso de las imágenes) como las temporales (en el caso de los datos de audio).

2.1.1 Fusión de Redes Neuronales Convolucionales en Bajo, Mediano y Alto Nivel

La fusión de Redes Neuronales Convolucionales es una técnica empleada para combinar diferentes modelos de CNNs con el objetivo de mejorar el rendimiento del sistema. Esta fusión puede realizarse en diferentes niveles: bajo, mediano y alto, cada uno con métodos y características particulares.

Fusión en bajo nivel

La fusión en bajo nivel de CNNs se refiere a la combinación de características de bajo nivel antes de que se realice cualquier procesamiento por las redes convolucionales. Este nivel de fusión suele implicar la concatenación o combinación de datos de entrada (Wang & Gupta, 2018).

Los datos se combinan antes de ser alimentados a la red convolucional por lo que a menudo requiere preprocesamiento para normalizar y alinear los datos, como la alineación de imágenes o la normalización de los valores de los píxeles (Han, et.al, 2015).

Fusión en mediano nivel

La fusión en mediano nivel implica la combinación de características intermedias extraídas de capas internas de diferentes CNNs. Este nivel de fusión utiliza la representación de características más abstractas que ya han sido procesadas parcialmente (He, et. al, 2016)

Combina características extraídas de capas convolucionales intermedias de diferentes redes, implicando el uso de múltiples redes CNN paralelas que extraen diferentes conjuntos de características (Huang, et.al, 2017)

Fusión en alto nivel

La fusión en alto nivel se refiere a la combinación de decisiones o salidas de diferentes CNNs. Este nivel de fusión se centra en integrar las predicciones finales para tomar una decisión más robusta (Ghosh, et. al, 2019)..

Combina las salidas o decisiones finales de múltiples CNNs según la confianza o la precisión de cada red.

2.1.2 Reconocimiento de Emociones en Expresiones Faciales bajo modelos de CNN

El reconocimiento de emociones a través de expresiones faciales implica el análisis de imágenes para identificar características faciales relevantes que correspondan a diferentes estados emocionales. Las CNN se entrenan utilizando grandes bases de datos de imágenes etiquetadas con diferentes emociones. Estas redes pueden aprender características como la curvatura de la boca, la posición de las cejas y la apertura de los ojos, que son indicativas de emociones como felicidad, tristeza, sorpresa y enojo (Trigeorgis, et.al, 2016) (He, et.al ,2016).

Antes de ser alimentadas a una CNN, las imágenes faciales suelen ser preprocesadas para mejorar la precisión del modelo. Esto puede incluir (Simonyan & Zisserman, 2014):

- Detección de Rostros: Usar algoritmos para localizar y extraer la región facial.
- Normalización: Ajustar el brillo y el contraste de las imágenes.
- Aumento de Datos: Aplicar transformaciones como rotaciones y escalados para generar más datos de entrenamiento y mejorar la robustez del modelo.

2.1.3 Reconocimiento de Emociones en Voz bajo modelos de CNN

El preprocesamiento de audios es esencial para mejorar la calidad de las señales acústicas y extraer características útiles para el análisis y reconocimiento de sonidos. Una técnica clave en este proceso es la extracción de los *Coeficientes Cepstrales en Frecuencia Mel* (MFCC). Para obtener los MFCC, se divide la señal de audio en pequeñas ventanas aplicando una ventana Hamming, luego se realiza una *Transformada de Fourier* (FFT) para convertir la señal al dominio frecuencial. Las frecuencias se pasan a través de filtros Mel, que simulan la percepción auditiva humana, y las energías resultantes se logarítmizan. Finalmente, se aplica una *Transformada Discreta del Coseno* (DCT) para obtener los coeficientes cepstrales. Estos coeficientes representan la envolvente espectral del audio y son muy útiles en el reconocimiento de patrones acústicos debido a su capacidad para capturar características relevantes y ser robustos ante variaciones en el ruido (Davis & Mermelstein,1980)

El reconocimiento de emociones en la voz implica el análisis de señales de audio para identificar patrones acústicos que correspondan a diferentes estados emocionales. Las CNN aplicadas a este ámbito a menudo utilizan espectrogramas, que son representaciones visuales de la frecuencia de la señal de audio a lo largo del tiempo, como entradas (Schuller, et.al, 2010).

Las señales de audio se convierten en espectrogramas mediante una transformación de Fourier. Estos espectrogramas se utilizan como entradas para las CNN,

que pueden aprender a reconocer patrones frecuenciales y temporales asociados con diferentes emociones.

El preprocesamiento del audio incluye:

- Eliminación de Ruido: Usar filtros para reducir el ruido de fondo.
- Normalización: Ajustar los niveles de volumen para que sean consistentes.
- Segmentación: Dividir el audio en segmentos más cortos si es necesario.

2.1.4 Sistemas Bimodales de Reconocimiento de Emociones con arquitecturas basadas en CNN

Las CNN bimodales integran tanto datos de imagen como de audio para mejorar la precisión y robustez del reconocimiento de emociones. Esta integración permite aprovechar las fortalezas de ambas modalidades y superar las limitaciones de los enfoques unimodales (Livingstone, et.al ,2018).

La integración de voz e imagen se logra mediante la combinación de características extraídas de ambas modalidades en una arquitectura de red neuronal conjunta. Esto puede implicar el uso de CNN para procesar imágenes faciales y espectrogramas de audio, seguido de la concatenación de las características extraídas y su procesamiento mediante capas adicionales de la red (Guo, et.al, 2016).

Una arquitectura típica de CNN bimodal puede incluir:

- Rama de Imagen: CNN para extraer características de imágenes faciales.
- Rama de Audio: CNN para extraer características de espectrogramas de audio.
- Fusión de Características: Concatenación o combinación de las características extraídas.
- Capas Completamente Conectadas: Clasificación final a partir de las características fusionadas.

2.2 Robot NAO

El robot NAO, mostrado en la Figura 10, desarrollado por la empresa francesa Aldebaran Robotics (ahora SoftBank Robotics), es un robot humanoide que ha sido ampliamente utilizado en diversos campos, incluyendo la educación, la investigación y el entretenimiento. Desde su lanzamiento en 2006, NAO ha destacado por su capacidad de interacción y su versatilidad, convirtiéndose en una herramienta clave para el estudio y desarrollo de la robótica social (SoftBank Robotics, n.d.).

Figura 10. Robot NAO

2.2.1 Robótica Humanoide

La robótica humanoide es una rama de la robótica que se enfoca en diseñar robots con forma y características humanas. Los robots humanoides, como NAO, son utilizados para investigar aspectos de la locomoción bípeda, la interacción humano-robot y la autonomía en entornos no estructurados. Estos robots son capaces de realizar movimientos y tareas similares a las de los humanos, lo que los hace ideales para la asistencia personal y la investigación en interacción social (Michaud, et.al, 2007).

2.2.2 Características Técnicas de NAO

NAO mide aproximadamente 58 centímetros de altura y pesa alrededor de 5.4 kilogramos. Está equipado con sensores, cámaras, micrófonos y altavoces que le permiten

percibir y reaccionar a su entorno. Su sistema operativo, NAOqi, facilita la programación y el control del robot, permitiendo a los desarrolladores crear comportamientos complejos a través de diversas plataformas de programación como Python, C++ y Java. Además, NAO cuenta con 25 grados de libertad, lo que le permite realizar movimientos fluidos y naturales (Gouaillier, et.al, 2009).

2.2.3 Aplicaciones de NAO

En el ámbito educativo, NAO se utiliza como una herramienta didáctica para enseñar programación y robótica a estudiantes de diferentes niveles. Su capacidad para interactuar de manera natural lo convierte en un recurso valioso para explorar conceptos de inteligencia artificial y aprendizaje automático. En la investigación, NAO ha sido utilizado para estudiar la interacción humano-robot, la comunicación no verbal y el desarrollo de algoritmos de control para robots humanoides. En el entretenimiento, NAO ha sido empleado en proyectos artísticos y performances debido a su capacidad de realizar movimientos coordinados y expresar emociones (Kudoh, et.al, 2005) (Moro, et.al, 2012).

El robot NAO ha tenido un impacto significativo en la investigación y la educación. Su accesibilidad y facilidad de programación han permitido a instituciones educativas y centros de investigación explorar nuevas fronteras en la robótica y la inteligencia artificial. Al proporcionar una plataforma versátil y adaptable, NAO ha fomentado la innovación y el desarrollo de nuevas aplicaciones en diversos campos.

2.3 Bases de datos emocionales

En el ámbito del reconocimiento de emociones, las bases de datos juegan un papel crucial. Proporcionan los datos necesarios para entrenar y evaluar modelos de aprendizaje automático y otras técnicas de inteligencia artificial. Entre las bases de datos más utilizadas se encuentran FER 2013 (*Facial Emotion Recognition* 2013) CK+ (Extended Cohn-Kanade Dataset) y RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song). Estas bases de datos son fundamentales para el desarrollo de sistemas

precisos y robustos que puedan identificar y clasificar emociones humanas de manera efectiva.

2.3.1 Base de Datos FER 2013

La base de datos FER2013 es un conjunto de datos ampliamente utilizado en la investigación de reconocimiento de expresiones faciales. Fue introducida en el contexto del Desafío de Reconocimiento de Expresiones Faciales (*Facial Expression Recognition Challenge*) en la Conferencia Internacional de Reconocimiento de Patrones (ICPR, *por sus siglas en inglés*) en 2013. Esta base de datos ha servido como un estándar para evaluar y comparar el rendimiento de diferentes algoritmos y modelos en el ámbito del reconocimiento de emociones (Li & Deng, 2020).

FER2013 consta de 35,887 imágenes de rostros etiquetadas, distribuidas en siete categorías de emociones: ira, disgusto, miedo, felicidad, tristeza, sorpresa y neutro. Las imágenes son de resolución 48x48 píxeles y están en escala de grises, lo que presenta un desafío adicional para los modelos de reconocimiento facial debido a la baja resolución y la falta de información de color.

Las imágenes están divididas en conjuntos de entrenamiento, validación y prueba.

- Conjunto de entrenamiento: 28,709 imágenes
- Conjunto de validación: 3,589 imágenes
- Conjunto de prueba: 3,589 imágenes

La Figura 11 muestra imágenes tomadas aleatoriamente de la base de datos FER 2013.

Figura 11. Imágenes contenidas en el dataset FER2013

2.3.1 Base de Datos CK+

La base de datos CK+ es una de las más populares y utilizadas en el ámbito del reconocimiento de emociones faciales. Se trata de una extensión de la base de datos original Cohn-Kanade, que se ha ampliado para incluir más sujetos y más expresiones emocionales. La base de datos original fue desarrollada en los años 2000 por Takeo Kanade y Jeffrey Cohn. La versión extendida, CK+, se lanzó en 2010, añadiendo más secuencias de vídeo y una mayor diversidad en las expresiones faciales (Lucey, et.al, 2010)

CK+ contiene 593 secuencias de vídeo de 123 sujetos, con cada secuencia capturando una transición desde una expresión neutral a una de las siete expresiones emocionales básicas: felicidad, tristeza, sorpresa, miedo, asco, enojo y desprecio. Además de los vídeos, también incluye anotaciones detalladas sobre los puntos de acción facial (Kanade, et.al, 2000)

Esta base de datos es ampliamente utilizada en la investigación de reconocimiento facial y en la validación de algoritmos de detección de emociones. Es especialmente útil para estudios que requieren datos de alta calidad y anotaciones precisas.

La Figura 12 muestra imágenes tomadas aleatoriamente de la base de datos CK+.

Figura 12. Imágenes tomadas de la base de datos CK+

2.3.1 Base de Datos RAVDESS

La base de datos RAVDESS es una colección de grabaciones de audio y vídeo que incluye expresiones emocionales de habla y canto. Es una herramienta valiosa para el estudio de emociones tanto a través de señales visuales como auditivas.

RAVDESS fue desarrollada por un equipo de investigadores de la Universidad de Ryerson en Canadá, y se publicó en 2018. Fue diseñada para superar las limitaciones de las bases de datos existentes al incluir una mayor variedad de emociones y modalidades de expresión, contiene 7356 archivos de audio y vídeo grabados por 24 actores profesionales. Las grabaciones incluyen 8 emociones (felicidad, tristeza, enojo, miedo, asco, sorpresa, calma y neutral) expresadas en discursos y canciones, con múltiples repeticiones y variaciones en intensidad emocional (Livingstone, et.al ,2018).

RAVDESS es utilizada para el desarrollo y evaluación de sistemas de reconocimiento de emociones multimodales, que analizan tanto las señales visuales como auditivas. Su riqueza en datos y su diversidad de emociones hacen que sea especialmente útil para estudios que buscan comprender la expresión emocional en diferentes contextos.

Capítulo 3

GENERACIÓN DE LA BASE DE DATOS EMO-MX-NAO CON EL ROBOT HUMANOIDE NAO

A lo largo de este capítulo se abordará la generación de la base de datos propia EMO-MX-NAO bajo la IHR con el robot humanoide NAO.

3.1 Club de teatro de la Universidad Tecnológica de la Mixteca

Para la generación de la base de datos emocional, se contó con el apoyo de los miembros del club de teatro, conformado por 9 integrantes: 5 hombres y 4 mujeres, con edades de 18 a 25 años, alumnos de la Universidad Tecnológica de la Mixteca que asisten a los talleres impartidos en el Centro de Actividades Culturales (CAC) de la misma institución educativa.

La creación del *dataset*, resultado de 12 horas de grabación segmentadas en tres sesiones de cuatro horas cada una dado el tiempo de descanso necesario para el correcto funcionamiento del robot, se dividió en dos partes, donde la primera contiene las expresiones faciales que los actores interpretan después la interacción con el robot NAO. Se obtuvo un conjunto de datos para expresiones faciales nombrado EMO-MX-EF-NAO creado utilizando la cámara superior del robot NAO y una cámara externa al mismo. EMO-MX-SP-NAO es el conjunto de datos de emociones en voz capturadas por un micrófono externo al robot.

3.1.1 EMO-MX-EF-NAO

La Figura 13 describe los pasos seguidos para la adquisición de las imágenes para la base de datos de expresiones faciales.

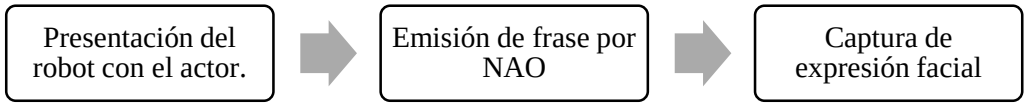


Figura 13. Flujo de captura de expresión facial

- Presentación del robot con el actor: NAO comienza la interacción presentando la actividad y describiendo el flujo de trabajo.
- Emisión de frase por NAO: Por cada emoción básica NAO pronunció un total de 10 frases. El orden de representación de las emociones fue: tristeza, miedo, aversión, ira, sorpresa y alegría.
- Al final de cada frase, fueron tomadas cuatro fotografías consecutivas: dos con la cámara superior del robot y dos con la cámara externa.

Las frases mostradas a continuación, son mensajes emitidos por el robot humanoide NAO que estimulan cada una de las seis emociones básicas.

- Alegría

    1. ¿Sabes cómo queda un mago después de comer? Queda magordito.
    2. ¿Sabes qué le dice un techo a otro? Techo de menos.
    3. ¿Sabes dónde cuelga súperman su súper capa? En su perchero.
    4. ¿Sabes para que va una caja al gimnasio? Para hacerse caja fuerte.
    5. ¿Sabes por qué se suicidó el libro de matemáticas? Porque tenía muchos problemas.
    6. Si los zombies se deshacen con el paso del tiempo, ¿son biodegradables?
    7. ¿Qué son 50 físicos y 50 químicos juntos? Pues cien tíficos.
    8. ¿Sabes cuál es el secreto de un moco? Ser viscoso y pegajoso.
    9. ¿Por qué le dio un paro cardiaco a la impresora? Parece que tuvo una impresión muy fuerte.
    10. ¿Por qué Bob Esponja no va al gimnasio? Porque ya está cuadrado.

- Aversión
    1. Revisa tus zapatos. Parece que has pisado la popó de un perro.
    2. El sandwich que me he comido parece que tenía algo podrido.
    3. Creo que me he lavado las manos con aguas negras
    4. Encontré una rata muerta en mi habitación.
    5. Esta mañana me he despegado cuatro mocos.
    6. No me he bañado en todo un mes.
    7. Encontré tres moscas en mi sopa de fideos.
    8. Creí que era chocolate, pero no. Era excremento de un ratón.
    9. En la oscuridad total, escuché un susurro gélido que me hizo temblar hasta los huesos.
    10. Las sombras danzaban, una oscuridad viva que devoraba la luz, y mi piel se erizaba.
- Ira
    1. Eres un tonto y flojo.
    2. Todo lo haces mal.
    3. Mi sistema se ha dañado por tu culpa.
    4. A este paso, jamás terminarás el proyecto.
    5. Deberías tomar un baño más a menudo. Hueles muy mal.
    6. ¡Llegar tarde es lo único que sabes hacer!
    7. Dejaré de trabajar contigo. ¡Sólo me usas a tu conveniencia!
    8. Podrías avanzar más rápido, pero no te gusta trabajar.
    9. Al paso que vas, tendrás obesidad en un par de meses.
    10. ¡Me he hecho daño por tu culpa!
- Miedo
    1. Cuentan que después de la una de la mañana, se asoma el espíritu de una chica que falleció cerca de COBAO.
    2. La nueva cepa de coronavirus es más contagiosa y letal.
    3. Huajuapan es el municipio con más infectados en la Mixteca Oaxaqueña
    4. He visto deambular a una mujer con vestido blanco todas las noches por estos rumbos.

5. Es posible que las reservas de agua potable se terminen en cincuenta años.
6. Debemos tener mucho cuidado. Los ladrones han logrado entrar a robar al menos a diez hogares en el fraccionamiento.
7. Me parece haber visto un par de serpientes en la entrada.
8. Creo que he sido abducido. No recuerdo muchas cosas desde hace un par de días.
9. En la oscuridad total, escuché un susurro gélido que me hizo temblar hasta los huesos.
10. Las sombras danzaban, una oscuridad viva que devoraba la luz, y mi piel se erizaba.

- Sorpresa
    1. ¡Gané la lotería!
    2. ¡Me iré de viaje a Japón!
    3. ¡Ganamos el primer lugar en el torneo de Fútbol!
    4. ¡He encontrado mil pesos tirados en la calle!
    5. ¡He adoptado un cachorrito!
    6. ¡Al fin he conseguido el videojuego que quería!
    7. ¡Me han dado un trabajo en el extranjero!
    8. ¡Me han ascendido de puesto!
    9. ¡He terminado de armar el rompecabezas de dos mil piezas!
    10. ¡Este verano de iré de vacaciones a Hawai!
- Tristeza
    1. Me han cancelado mi visa y es posible que no vea a mi familia en dos años.
    2. Rechazaron mi artículo en el congreso.
    3. Mi perrito se ha perdido. Lo he buscado sin descanso los últimos tres días y no logro encontrarlo.
    4. Envenenaron a mi gato y no logré llegar a tiempo al veterinario.
    5. Mi mejor amigo se ha enfermado de coronavirus. Está muy grave en el hospital.
    6. Me han expulsado del equipo de futbol.

7. Ella me ha abandonado.
8. Sin mi perrito, no le encuentro sentido a la vida.
9. Me he quedado sin empleo y ahora no sé qué hacer
10. Me siento muy solo y ahora no tengo nadie con quien platicar.

La generación de la base de datos se efectuó bajo un ambiente controlado y con luz artificial, como se observa en la Figura 14, a fin de minimizar las variables que puedan afectar el aprendizaje del sistema de reconocimiento de expresiones faciales,

Figura 14. Interacción NAO-Actor

La Figura 15 contiene una muestra de las fotografías capturadas por la cámara del robot. Cada imagen original tiene un tamaño de 640 x 480 pixeles.

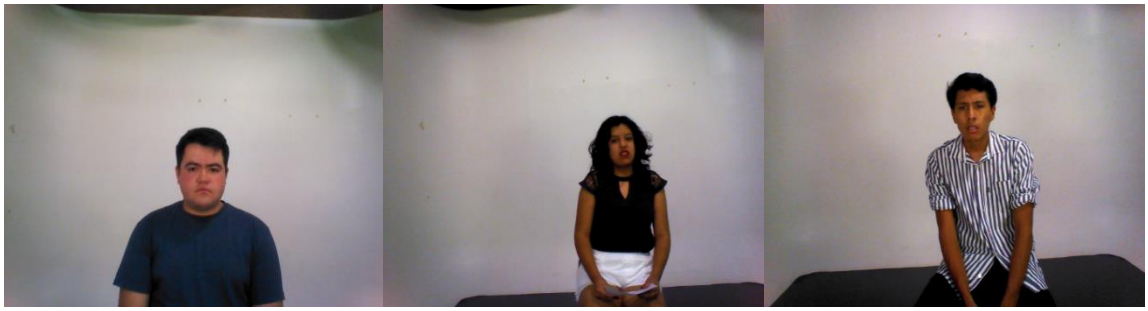


Figura 15. Emociones capturadas por la cámara del robot NAO

La Figura 16 contiene una muestra de las emociones capturadas por la cámara externa al robot. Cada imagen original tiene un tamaño de 5184 x 3456 pixeles.

Figura 16. Emociones capturadas por una cámara externa al robot NAO

La base de datos EMO-MX-EF-NAO contiene un total de 1236 imágenes de expresiones faciales para las seis emociones básicas en blanco y negro con un tamaño 192 x 192 pixeles, donde se ha recortado únicamente el rostro a fin de eliminar información no relevante. La Tabla 1 muestra la distribución de la base de datos.

Tabla 1. Distribución de imágenes en la base de datos EMO-MX-EF-NAO

Emoción	*Entrenamiento*	*Validación*	*Pruebas*
01- Enojo	145	31	30
02- Aversión	145	31	30
03- Miedo	145	31	30
04- Alegría	145	31	30
05- Tristeza	145	31	30
06- Sorpresa	145	31	30

La Figura 17 ilustra algunas imágenes contenidas en la base de datos EMO-MX-EF-NAO bajo diferentes emociones.

Figura 17. Imágenes contenidas en el dataset EMO-MX-EF-NAO

3.1.3 EMO-MX-SP-NAO

EMO-MX-SP-NAO representa la base de datos de audios generada a partir de la interacción con el robot NAO. A continuación, se enlistan las doce frases empleadas para esta base de datos. El conjunto de audios incluye cada frase grabada seis veces, una por cada emoción básica:

1. Son las once en punto.
2. Eso es exactamente lo que pasó.
3. Estoy de camino a la reunión.
4. Me pregunto de qué se trata esto.
5. El avión está casi lleno.
6. Quizás mañana haga frío.
7. Compraré un nuevo despertador.
8. Tengo una cita con el médico.
9. No olvides una chaqueta.
10. No ha llovido en varios días.
11. La superficie está resbaladiza.

12. Pararemos en un par de minutos.

La base de datos contiene en total 546 archivos. Los audios fueron grabados utilizando un micrófono externo al robot, bajo una frecuencia de muestreo de 48kHz, con una duración de 5 segundos cada uno, añadiendo intervalos de silencio al inicio y al final de la interpretación del actor.

La Tabla 2 muestra la distribución de los audios en la base de datos EMO-MX-SP-NAO

Tabla 2. Distribución de audios en la base de datos EMO-MX-SP-NAO

Emoción	*Entrenamiento*	*Validación*	*Pruebas*
01- Enojo	68	12	11
02- Aversión	68	12	11
03- Miedo	68	12	11
04- Alegría	68	12	11
05- Tristeza	68	12	11
06- Sorpresa	68	12	11

La base de datos EMO-MX-NAO se encuentra disponible para ser revisada en su totalidad en el siguiente enlace de GitHub:

https://github.com/CristellTL/EMO-MX-NAO

Capítulo 4

METODOLOGÍA

Como se presentó en el Capítulo 1, el sistema multimodal de reconocimiento de emociones propuesto efectúa el reconocimiento en dos señales: imagen y voz. A fin de comparar el rendimiento del sistema multimodal, debe evaluarse el comportamiento de la red neuronal como sistema monomodal a partir de cada una de las dos señales mencionadas.

En este capítulo se presenta el desarrollo de los sistemas monomodales a partir de una base de datos existente y la base de datos propia EMO-MX-NAO.

4.1 Sistema Reconocimiento de Expresiones Faciales

A continuación, se describen los distintos experimentos efectuados empleando dos arquitecturas diferentes de CNN para el reconocimiento de expresiones faciales con las bases de datos existentes FER2013 y CK+, además de las pruebas realizadas con la base de datos propia EMO-NAO-EF.

4.1.1 Experimento 1 CK+_3CC

El reconocimiento de emociones en imagen, se llevó a cabo inicialmente con la implementación de una red neuronal de 3 bloques de capas convolucionales seguidas de una capa de Max Pooling donde *padding = same* y dos capas completamente conectadas donde la última representa la capa de clasificación con seis clases al final de la red, para imágenes de 48x48 pixeles tal como se muestra en la Figura 18. Las pruebas del modelo se ejecutaron con el *dataset* CK+ descartando la emoción desprecio, puesto que se trata de una emoción compuesta por ira y aversión, y segmentando aleatoriamente el conjunto de datos en tres directorios siguiendo la regla de 70-15-15, 70% de los datos para entrenamiento, 15% para validación y 15% para *test.*

La distribución de los datos se muestra en la Tabla 3.

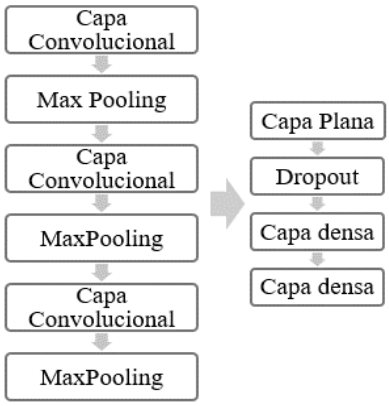


Figura 18. Primer modelo implementado para el reconocimiento de expresiones faciales

A continuación, se describen los distintos experimentos efectuados empleando dos arquitecturas diferentes de CNN para el reconocimiento de expresiones faciales con las bases de datos existentes FER2013 y CK+, además de las pruebas realizadas con la base de datos propia EMO-NAO-EF

- *Rotation_range:* 40
- *Width_shift_range:* 0.3
- *Height_shift_range:* 0.3
- *Zoom_range:* 0.3
- *Horizontal_flip:* True
- *Vertical_flip:* True

Tabla 3. Distribución de imágenes de la base de datos CK+

Emoción	Imágenes de entrenamiento	Imágenes de validación	Imágenes de prueba
01-Enojo	95	20	20
02-Aversión	125	26	26
03-Miedo	53	11	11
04-Alegría	145	31	31
05-Tristeza	60	12	12
06-Sorpresa	175	37	37

La Figura 19 muestra los resultados obtenidos en el primer experimento para 1000 *epochs* y un tamaño de lote de 16. Se observa la curva de precisión para entrenamiento y validación con resultados de 95.28% y 71.07% respectivamente, se muestra además una precisión de 71.01% en datos de prueba. Sin embargo, al analizar la matriz de confusión se observa un desbalance al clasificar mejor las emociones felicidad y sorpresa y un muy bajo reconocimiento en las clases miedo y tristeza.

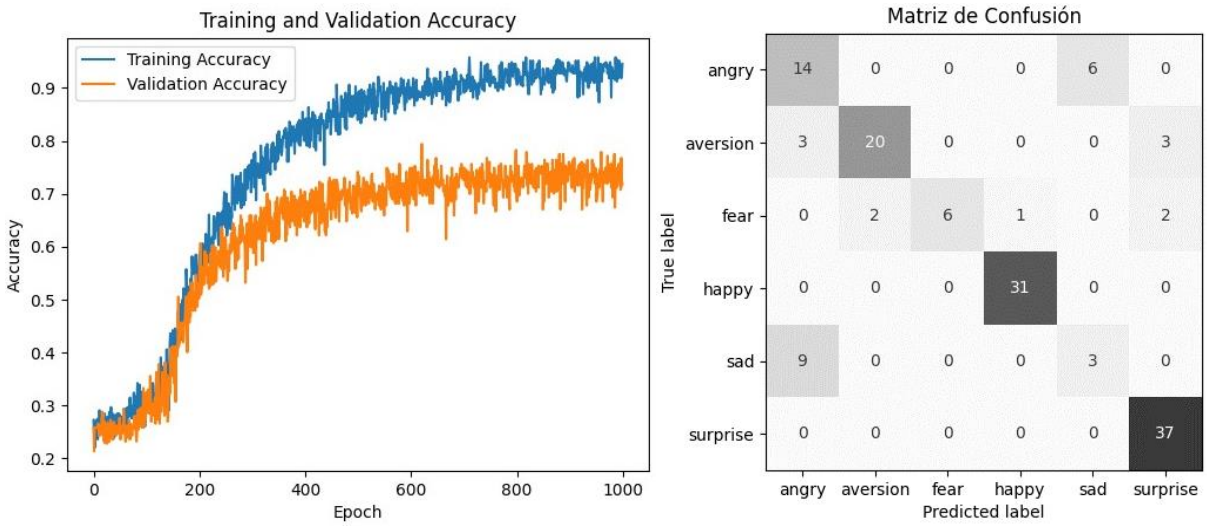


Figura 19. Resultados Experimento 1. Reconocimiento de Expresiones Faciales

4.1.2 Experimento 2 DB3_3CC

A fin de solucionar el problema de desbalance de la base de datos, es decir, igualar el número de imágenes por clase y generalizar la red neuronal, se optó por fusionar las bases de datos CK+ y FER2013, descartando nuevamente la emoción desprecio para la clasificación de las seis emociones básicas. La Figura 20 muestra el flujo de preprocesamiento de las imágenes antes de ser ingresadas a la red neuronal.

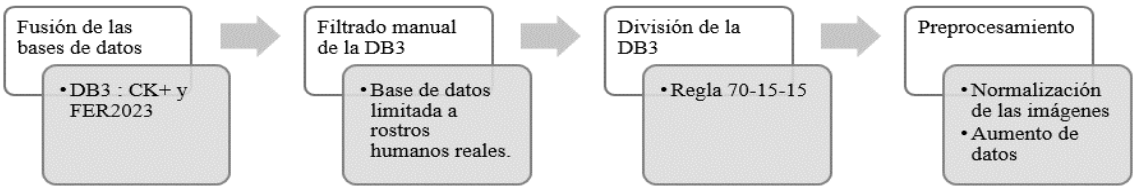


Figura 20. Preprocesamiento de la base de datos BD3

El resultado de la fusión inicial de las bases de datos da como resultado una tercera base de datos con un total de 36814 imágenes. Para obtener una comparación válida con la base de datos propia, se limitó la BD3 a imágenes de rostros humanos reales de entre 18 y 30 años de edad. Fueron descartadas también, imágenes con rostros de perfil y rostros incompletos. La Figura 21 muestra algunas de las imágenes descartadas de la BD3.

Figura 21. Muestra de imágenes descartadas de DB3

La Tabla 4, detalla la distribución de la base de datos final completamente balanceada con 3600 imágenes. El preprocesamiento de las imágenes antes de ingresar a la CNN siguió las mismas características del experimento 1.

Tabla 4. Distribución de imágenes en la DB3

Emoción	Imágenes de entrenamiento	Imágenes de validación	Imágenes de prueba
Enojo	400	100	100
Aversión	400	100	100
Miedo	400	100	100
Alegría	400	100	100
Tristeza	400	100	100
Sorpresa	400	100	100

La Figura 22, ilustra las curvas de precisión para las fases de entrenamiento y validación, así como la matriz de confusión de la arquitectura descrita en la Figura 18 considerando 1000 *epochs* y un tamaño de lote igual a 16.

Se observa una precisión de 59% en entrenamiento, 56% para validación y 59% en los datos de prueba, lo que indica un modelo limitado para el reconocimiento de expresiones faciales. Los resultados sugieren la necesidad de un incremento de capas

convolucionales que incrementen la complejidad de la red neuronal para la extracción de características y mejora en la precisión.

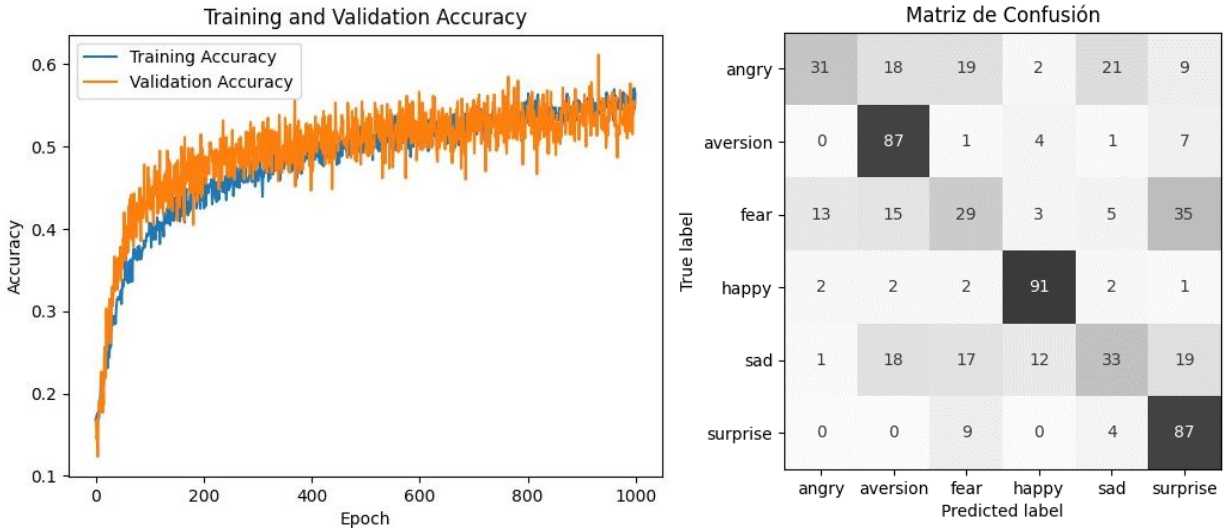


Figura 22. Resultados. Experimento 2

4.1.3 Experimento 3 BD3_5CC

La Figura 23 muestra la nueva arquitectura propuesta de manera eurística para el modelo unimodal de reconocimiento de emociones. Las capas convolucionales y de Max Pooling incluyen *padding = same* convenientemente para incluir un mayor número de capas en la red.

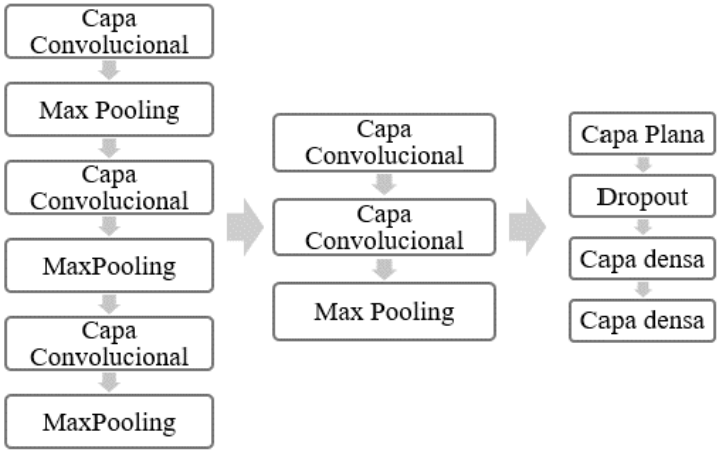


Figura 23. Segundo modelo implementado para el reconocimiento de expresiones faciales

La Tabla 5 contiene las variaciones de *epochs* y *batch_size* bajo un *learning_rate=0.001* configurados para diferentes entrenamientos utilizando la base de datos DB3 además de las precisiones obtenidas en las fases de entrenamiento, validación y prueba.

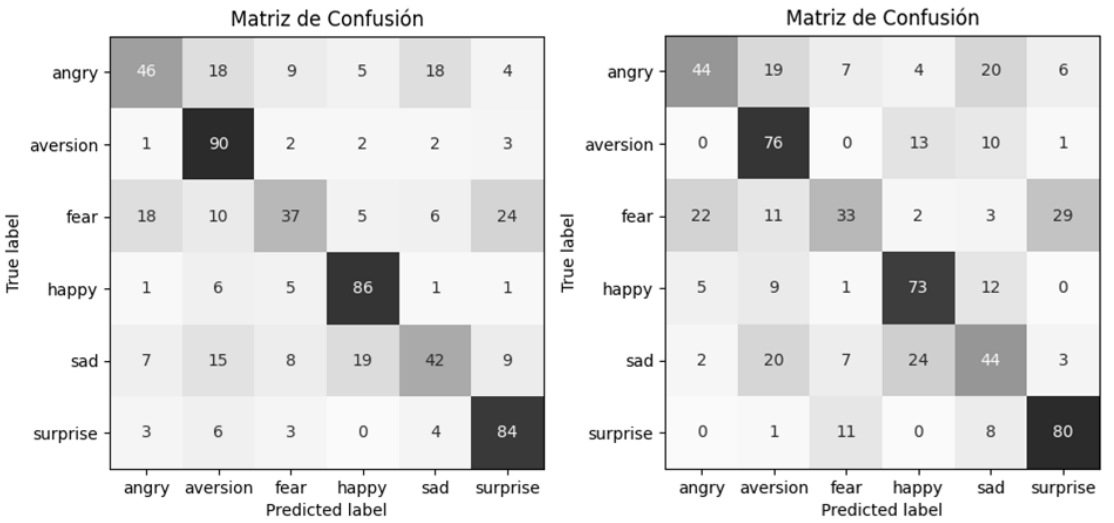
Tabla 5. Resultados obtenidos de la arquitectura 2 bajo diferentes hiperparámetros

Nombre del modelo	*Epochs*	*Batch size*	*Training accuracy*	*Validation accuracy*	*Test accuracy*
CNN_EF_5CC_16_800	800	16	0.62	0.60	0.64
CNN_EF_5CC_64_500	500	64	0.58	0.57	0.58
CNN_EF_5CC_128_2000	2000	128	0.65	0.62	0.64
CNN_EF_5CC_256_2000	2000	256	0.67	0.64	0.66

El modelo mejora a medida que el tamaño de batch incrementa, sin embargo, no se realizan prueban con lotes más grandes dado los limitados recursos computacionales bajo los que se efectuó cada prueba. La Figura 24 muestra la matriz de confusión de cada uno de los modelos descritos en la Tabla 5.

El modelo seleccionado para la clasificación de expresiones faciales es el CNN_EF_5CC_256_2000, con el que se obtienen mejores porcentajes de precisión para imágenes no vistas en la fase de entrenamiento.

Las emociones que son más susceptibles a ser clasificadas únicamente analizando la expresión facial son: aversión, felicidad y sorpresa, mientras enojo, miedo y tristeza requieren de mayor información para ser clasificadas correctamente.



a) CNN_EF_5CC_16_800 b) CNN_EF_5CC_64_500

c) CNN_EF_5CC_128_2000 d) CNN_EF_5CC__256_2000

Figura 24. Matrices de confusión obtenidas a partir de los modelos de la Tabla 5

4.1.4 Experimento 4 EMO-MX-EF-NAO

En esta sección, se muestra la evaluación de las dos arquitecturas de CNN para la clasificación de expresiones faciales descritas en los experimentos 1 y 3, Figuras 18 y 23 respectivamente a fin de comparar su desempeño con la base de datos EMO-MX-EF-NAO.

La Tabla 6 contiene los parámetros bajo los cuales se efectuaron los diferentes entrenamientos siguiendo la arquitectura mostrada en la Figura 18, una CNN con tres capas convolucionales y dos capas completamente conectadas bajo un *learning_rate=0.001* utilizando la base de datos EMO-MX-EF-NAO además de las precisiones obtenidas en las fases de entrenamiento, validación y prueba.

Tabla 6. Resultados obtenidos para 3CC en EMO-MX-EF-NAO

Nombre del modelo	*Epochs*	*Batch size*	*Training accuracy*	*Validation accuracy*	*Test accuracy*
CNN_EF_NAO_3CC_16_100	100	16	0.89	0.84	0.85
CNN_EF_NAO_3CC_32_100	100	32	0.90	0.86	0.86
CNN_EF_NAO_3CC_64_120	120	64	0.93	0.87	0.87
CNN_EF_NAO_3CC_128_100	100	128	0.92	0.85	0.86

Se observa, de forma general, una mejora en el desempeño de la red entrenada con la base datos propia, obtenida por el incremento en la calidad de las imágenes. En el mismo sentido, las matrices de confusión de los modelos listados en la Tabla 6 indican un balance en *test accuracy* al clasificar correctamente las seis emociones con porcentajes de precisión muy cercanos, tal como se observa en la Figura 25.

La Tabla 7 muestra las precisiones obtenidas en las etapas de entrenamiento, validación y pruebas de la arquitectura mostrada en la Figura 18, una CNN con cinco capas convolucionales y dos capas completamente conectadas. A fin de efectuar una correcta comparación, se utilizó *learning rate= 0.001*.

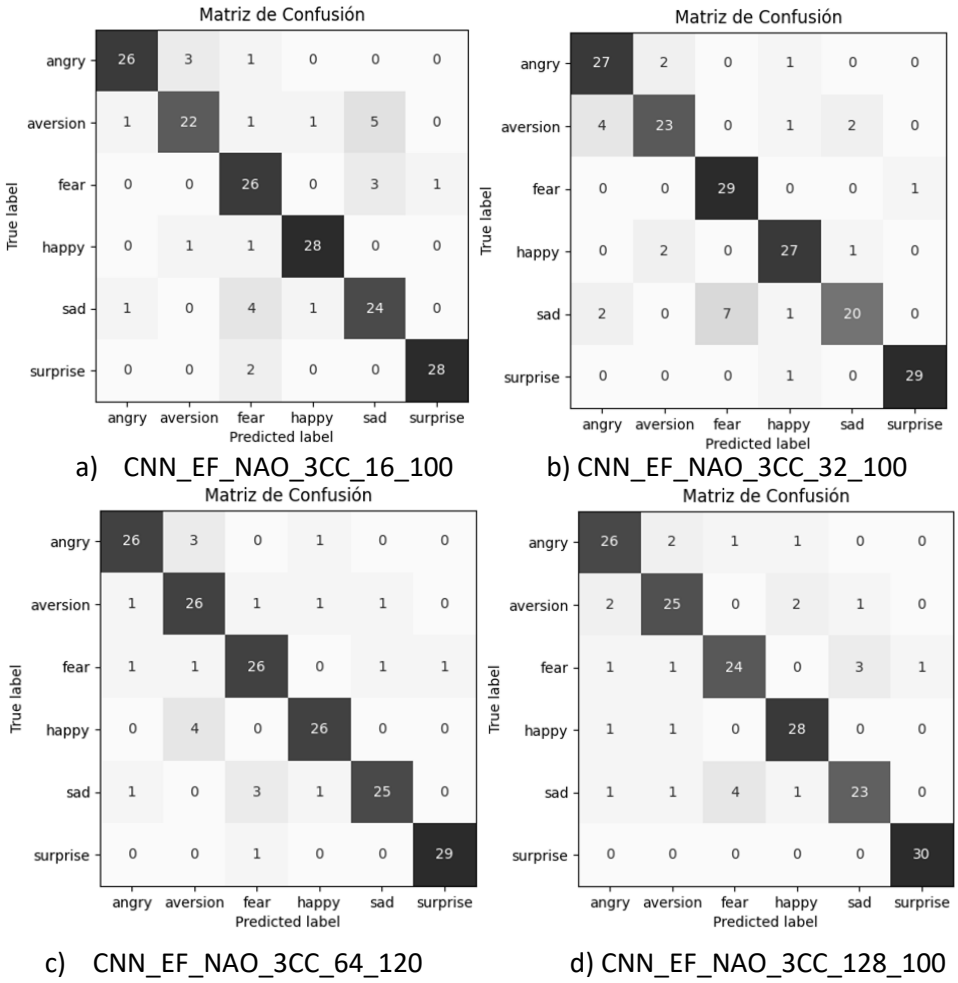


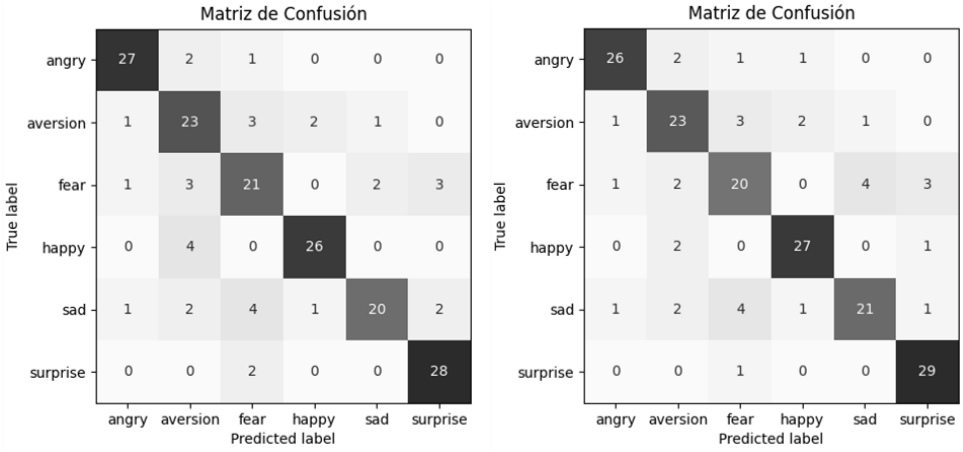
Figura 25. Matrices de confusión de los modelos mostrados en la Tabla 6

Tabla 7. Resultados obtenidos para 5CC en EMO-MX-EF-NAO

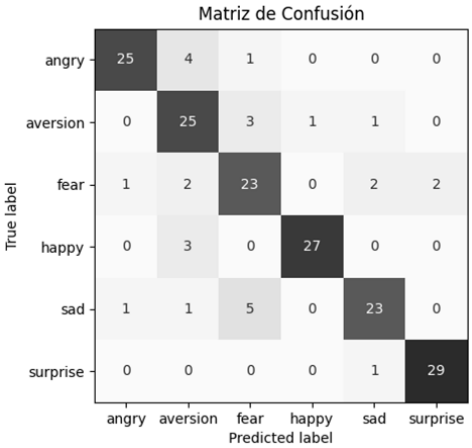
Nombre del modelo	*Epochs*	*Batch size*	*Training accuracy*	*Validation accuracy*	*Test accuracy*
CNN_EF_NAO_5CC_16_100	100	16	0.81	0.81	0.80
CNN_EF_NAO_5CC_64_120	120	64	0.86	0.82	0.81
CNN_EF_NAO_5CC_128_80	80	128	0.88	0.84	0.84

La Figura 26 muestra las matrices de confusión de los modelos descritos en la Tabla 7. Se observa que la arquitectura con cinco capas convolucionales sigue teniendo

un desempeño similar en las seis clases lo que indican un balance en la predicción, sin embargo, los porcentajes de precisión son menores que los obtenidos en la arquitectura con tres capas convolucionales.



a)CNN_EF_NAO_5CC_16_100 b) CNN_EF_NAO_5CC_32_500



c)CNN_EF_NAO_5CC_128_80

Figura 26. Matrices de confusión de los modelos mostrados en la Tabla 7

Finalmente, el modelo seleccionado para la clasificación de expresiones faciales es CNN_EF_NAO_3CC_64_120 dado los altos porcentajes de precisión que presenta y el balance en la predicción.

4.2 Sistema Reconocimiento de Emociones en Voz

A continuación, se describen los distintos experimentos efectuados empleando dos arquitecturas diferentes de CNN para el reconocimiento de emociones en voz con las bases de datos RAVDESS y EMO-MX-SP-NAO

4.2.1 Experimento 1 RAVDESS_3CNN

Para la clasificación de las emociones mediante la voz, también se ha implementado una red CNN, con la arquitectura que se muestra en la Figura 27.

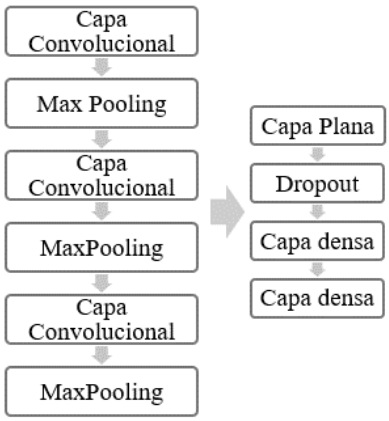


Figura 27. Arquitectura heurística de CNN para la clasificación de emociones en habla

Para entrenar y evaluar el modelo, se utilizó el subconjunto de datos *emotional speech* de la base de datos RAVDESS que cuenta con 1146 audios. Los datos fueron distribuidos tal como se muestra en la Tabla 8 de acuerdo a la regla 70-15-15.

Tabla 8. Distribución de audios en la base de datos RAVDESS

Emoción	Audios de entrenamiento	Audios de validación	Audios de prueba
Enojo	131	30	30
Aversión	131	30	30
Miedo	131	30	30
Alegría	131	30	30
Tristeza	131	30	30
Sorpresa	131	30	30

Cada archivo de audio, pasó por una transformación a espectrogramas utilizando funciones de la librería *librosa* bajo los siguientes parámetros para poder ingresar a la CNN:

- Ventana de análisis (n_fft): 2048 muestras.
- Solapamiento (hop_length): 512 muestras, 75%.
- Tipo de ventana (window): Ventana de Hann.
- Centro de la ventana (center): True.
- Modo de borde (pad_mode): reflect

La Figura 28 contiene una muestra de espectrogramas generados a partir de la base de datos RAVDESS.

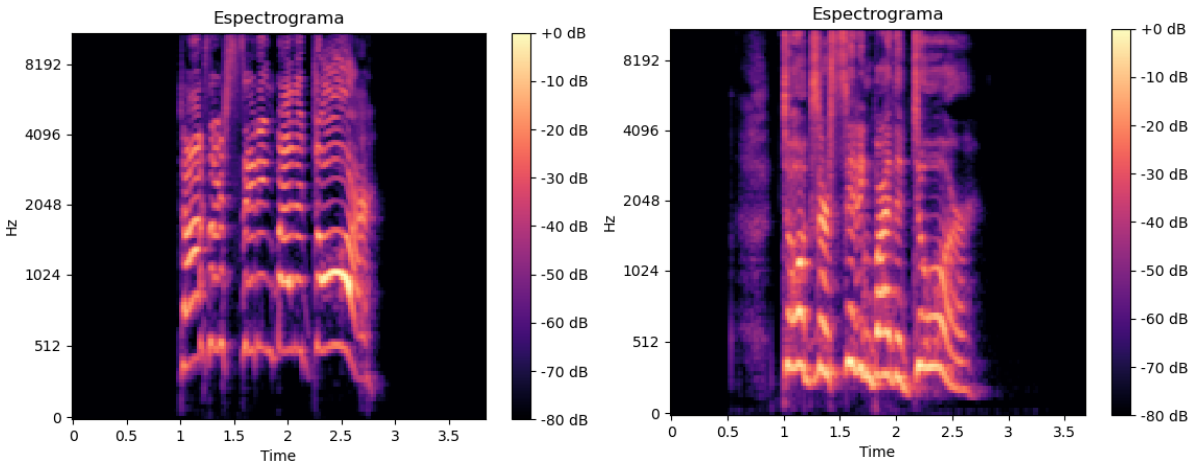


Figura 28. Espectrogramas generados a partir de la base de datos RAVDESS

Los espectrogramas fueron recortados y redimensionados a 48x48 pixeles para conservar únicamente las magnitudes de las componentes de frecuencia en relación al tiempo que puedan proporcionar información útil, por lo que se eliminó la descripción de color, el título de la figura y la descripción de los ejes.

La Figura 29 muestra algunas de las imágenes que ingresan a la CNN de clasificación de emociones en voz.

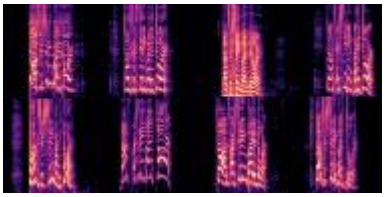


Figura 29. Muestra de imágenes que ingresan a la CNN de reconocimiento en voz

La Tabla 9 muestra los parámetros bajo los cuales se efectuaron diferentes entrenamientos de la CNN de clasificación de emociones en voz con la arquitectura descrita en la Figura 27 y bajo *learning_rate= 0.001*

Tabla 9. Resultados obtenidos para la clasificación de emociones en voz para la primera arquitectura

Nombre del modelo	*Epochs*	*Batch size*	*Training accuracy*	*Validation accuracy*	*Test accuracy*
CNN3_Voice_RAV_16_30	30	16	0.76	0.62	0.58
CNN3_Voice_RAV_32_25	25	32	0.76	0.66	0.61
CNN3_Voice_RAV_64_20	20	64	0.73	0.64	0.59

La CNN de sólo tres capas convolucionales para el reconocimiento de emociones en voz mostró, generalmente, mejores resultados en términos de precisión en comparación con el modelo de CNN de 3 capas convolucionales de reconocimiento de expresiones faciales. Además, tal como se muestra en la matriz de confusión de la Figura 30, las emociones de enojo, aversión y miedo se clasifican con una precisión superior al 80% mostrando ser el complemento de la clasificación de expresiones faciales para el reconocimiento de emociones. Tristeza sigue siendo la emoción con menor desempeño dado que presenta un porcentaje de precisión del 30% tanto en imagen como en voz.

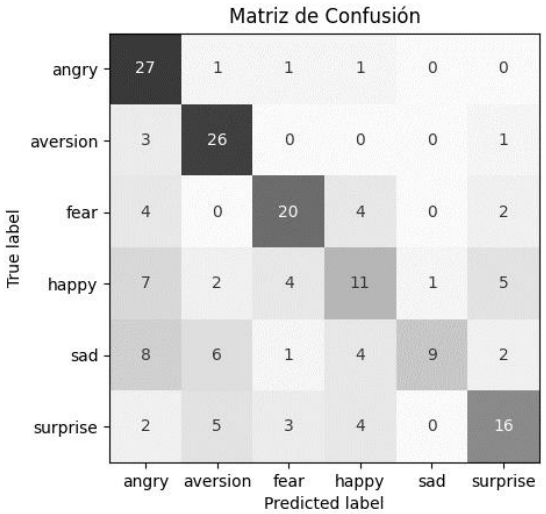


Figura 30 Matriz de confusión para el modelo CNN3_Voice_RAV_32_25

En el próximo experimento, se efectúa entrenamientos con una CNN de mayor complejidad a fin de evaluar si el desempeño de los modelos mejora con un mayor número de capas convolucionales tal como sucedió con los modelos de CNN para expresiones faciales.

4.2.2 Experimento 2 RAVDESS_6CNN

La segunda arquitectura propuesta para el reconocimiento de emociones en voz consta de seis capas convolucionales y dos capas completamente conectadas, donde la última corresponde a la fase de clasificación, como se observa en la Figura 31.

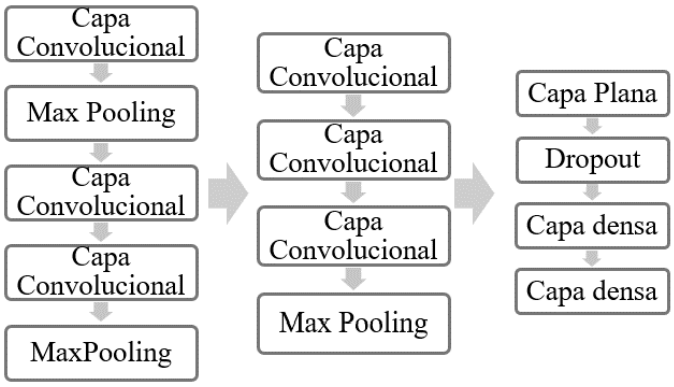


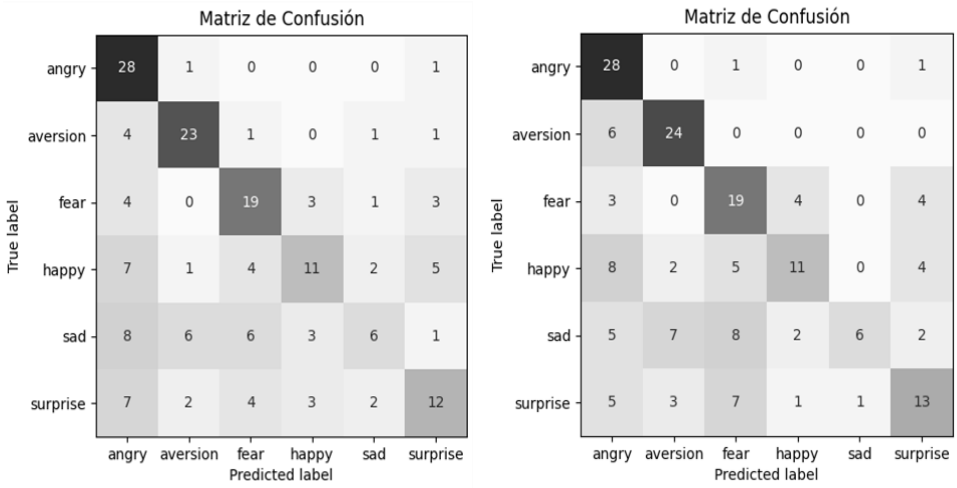
Figura 31. Segundo modelo heurístico propuesto para el reconocimiento de emociones en voz

La Tabla 10 contiene las variaciones de *epochs* y *batch_size* bajo un *learning_rate=0.001* configurados para diferentes entrenamientos utilizando la base de datos RAVDESS además de las precisiones obtenidas en las fases de entrenamiento, validación y prueba.

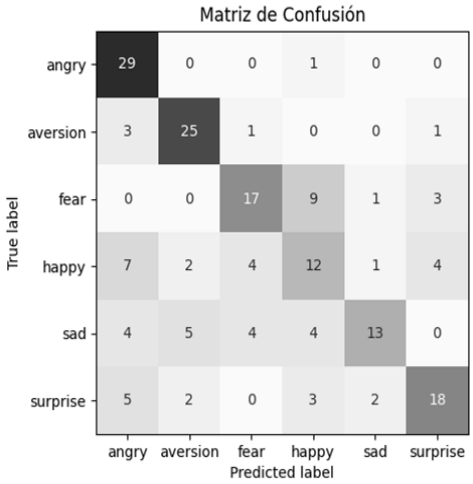
Tabla 10. Resultados obtenidos para la clasificación de emociones en voz para la segunda arquitectura

Nombre del modelo	*Epochs*	*Batch size*	*Training accuracy*	*Validation accuracy*	*Test accuracy*
CNN6_Voice_RAV_16_25	25	16	0.81	0.61	0.55
CNN6_Voice_RAV_32_20	20	32	0.71	0.60	0.56
CNN6_Voice_RAV_64_25	25	64	0.85	0.64	0.63

En la Figura 32 las matrices de confusión indican, para el modelo de seis capas convolucionales, una mejor clasificación en las emociones: enojo, aversión y sorpresa.



a) CNN6_Voice_RAV_16_25 b) CNN6_Voice_RAV_32_20



a) CNN6_Voice_RAV_64_25

Figura 32. Matrices de confusión de los modelos descritos en la Tabla 10

Se observa que, en el caso de las emociones en voz, se obtienen precisiones más altas con modelos de CNN con un menor número de capas convolucionales, lo que puede ser resultado del tamaño del conjunto de datos.

El modelo seleccionado para la clasificación de emociones en habla es el CNN6_Voice_RAV_64_25, con el que se obtienen mejores porcentajes de precisión para imágenes no vistas en la fase de entrenamiento.

4.2.3 Experimento 3 EMO-MX-SP-NAO

El conjunto de datos EMO-MX-SP-NAO se empleó para los entrenamientos de las arquitecturas descritas en las Figuras 27 y 31 con tres y seis capas convolucionales respectivamente.

A partir de cada uno de los 546 audios del *dataset* se obtuvo los espectrogramas con un tamaño de 192x192 píxeles que ingresarán a la red neuronal.

La Figura 33 contiene una muestra de los espectrogramas del *dataset* EMO-MX-SP-NAO.

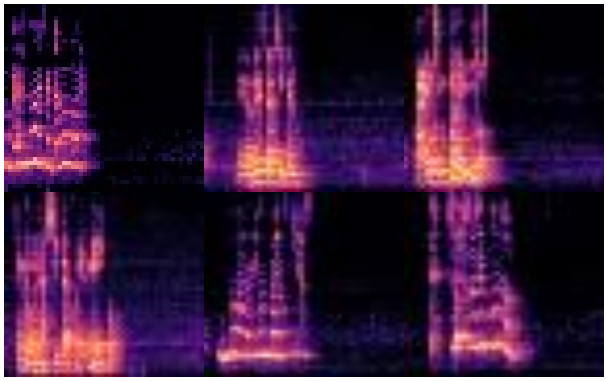


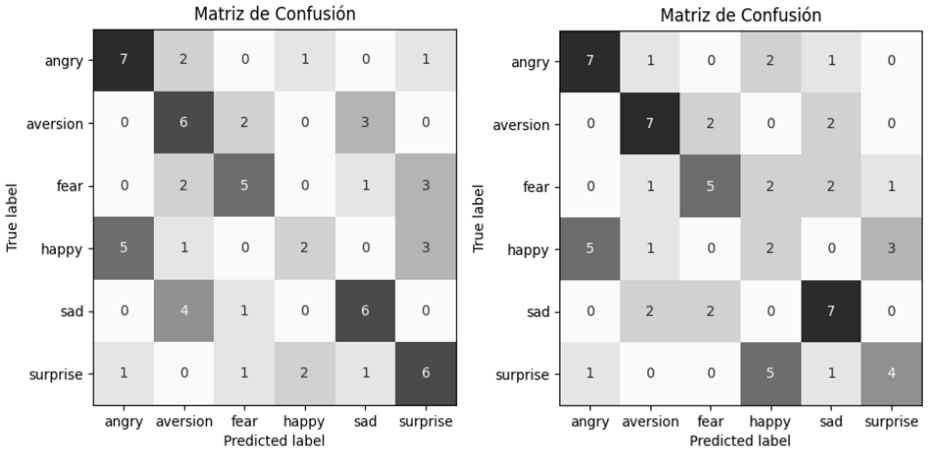
Figura 33. Muestra de los espectrogramas de la base de datos EMO-MX-SP-NAO

La Tabla 11 muestra los parámetros bajo los cuales se efectuaron diferentes entrenamientos de la CNN de clasificación de emociones en voz con la arquitectura descrita en la Figura 27 y *learning_rate= 0.001*.

Tabla 11. Resultados obtenidos con la base de datos propia y 3CC para clasificación de emociones en voz

Nombre del modelo	*Epochs*	*Batch size*	*Training accuracy*	*Val accuracy*	*Test accuracy*
CNN3_Voice_NAO_16_45	45	16	0.59	0.51	0.48
CNN3_Voice_NAO_32_45	45	32	0.58	0.54	0.48
CNN3_Voice_NAO_64_45	45	64	0.51	0.50	0.46

En la arquitectura de red de tres capas convolucionales, los porcentajes de precisión son muy bajos en el reconocimiento de emociones en voz. Las matrices de confusión mostradas en la Figura 34 para las pruebas de los modelos entrenados en la Tabla 11 muestran alta dispersión en la clasificación.



a) CNN3_Voice_NAO_16_45 b) CNN3_Voice_NAO_32_45

c) CNN3_Voice_RAV_64_45

Figura 34. Matrices de confusión de los modelos descritos en la Tabla 11

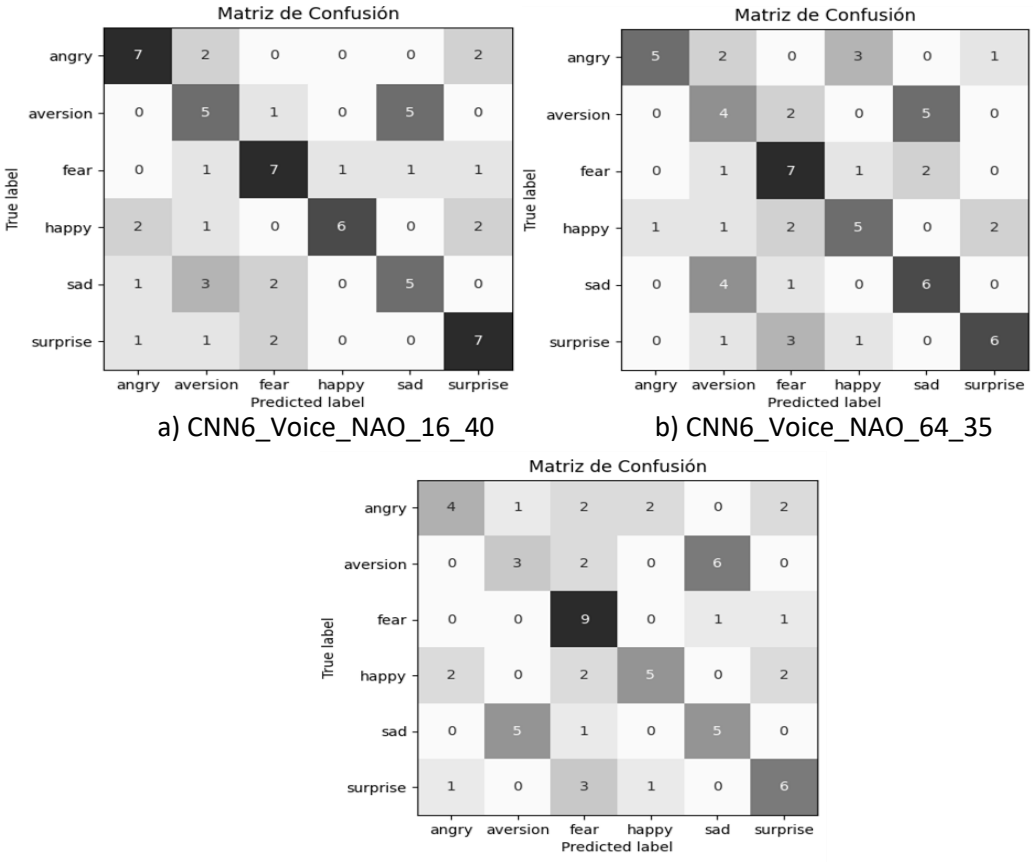
La Tabla 12 muestra el desempeño de la CNN cuya arquitectura es descrita en la Figura 31 bajo diferentes parámetros y con *learning_rate = 0.001*.

La clasificación de emociones en voz muestra un menor sobre ajuste de los modelos para la red de seis capas convolucionales que para la red de tres capas convolucionales mientras que la precisión en las imágenes de prueba incrementa ligeramente.

Tabla 12. Resultados obtenidos con la base de datos propia y 6CC para clasificación de emociones en voz.

Nombre del modelo	*Epochs*	*Batch_size*	*Training accuracy*	*Validation accuracy*	*Test accuracy*
CNN6_Voice_NAO_16_40	40	16	0.77	0.52	0.56
CNN6_Voice_NAO_64_35	35	32	0.74	0.56	0.50
CNN6_Voice_NAO_32_40	45	64	0.72	0.50	0.50

La Figura 35, muestra las matrices de confusión de los modelos de la Tabla 12. Se ha seleccionado el modelo CNN6_Voice_NAO_16_40 dado su desempeño en validación y prueba y un mayor balance en la clasificación de emociones.



a) CNN6_Voice_NAO_16_40

b) CNN6_Voice_NAO_64_35

c) CNN3_Voice_NAO_32_40

Figura 35. Matrices de confusión para los modelos de la Tabla 12

4.3 Sistema Multimodal de Reconocimiento de Emociones

Los siguientes experimentos se han realizado con los tres niveles de fusión de CNN: bajo, medio y alto utilizando las bases de datos preexistentes y la base de datos EMO-MX-NAO.

4.3.1 Fusión en bajo nivel

Para la fusión en bajo nivel, se optó por utilizar la arquitectura de CNN de la Figura 36, una red neuronal convolucional de tres capas convolucionales y dos capas completamente conectadas.

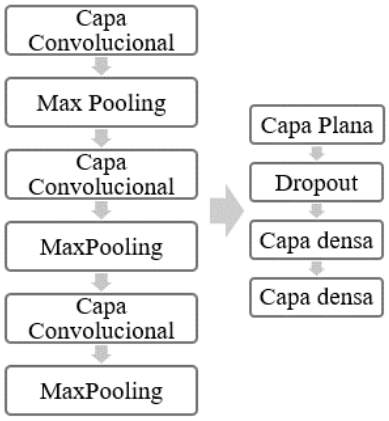


Figura 36. Arquitectura de CNN para el modelo de fusión a bajo nivel

La fusión en bajo nivel implica combinar todas las características existentes antes de ingresar a la red neuronal por lo que se unió aleatoriamente cada espectrograma con una expresión facial de la misma clase.

Fusión en BD3 y RAVDESS

La Figura 37 ilustra el resultado de la unión del espectrograma-expresión facial en una imagen de 96x48 pixeles.

Figura 37. Combinación de características a bajo nivel

En el *dataset* resultante de la unión de la BD3 con RAVDESS se descartó aquellas expresiones faciales que no contaban con un espectrograma que se le pudiera asignar. La distribución del conjunto de datos empleado se muestra en la Tabla 13.

Tabla 13. Distribución de imágenes para la base de datos en fusión a bajo nivel

Emoción	Imágenes de entrenamiento	Imágenes de validación	Imágenes de prueba
Enojo	131	30	30
Aversión	131	30	30
Miedo	131	30	30
Alegría	131	30	30
Tristeza	131	30	30
Sorpresa	131	30	30

La Tabla 14 muestra los resultados de diferentes entrenamientos para el sistema multimodal de reconocimiento de emociones con fusión de bajo nivel aplicado a las bases de datos BD3 y RAVDESS.

Tabla 14.Modelos de fusión de bajo nivel BD3 y RAVDESS

Nombre del modelo	*Epochs*	*Batch_size*	*Training accuracy*	*Validation accuracy*	*Test accuracy*
MLLF_3CC_16_13	13	16	0.75	0.72	0.66
MLLF_3CC_32_18	18	32	0.72	0.68	0.64
MLLF_3CC_64_15	15	64	0.78	0.72	0.64

Las matrices de confusión mostradas en la Figura 38, muestran el comportamiento de los modelos descritos en la Tabla 14 con imágenes de prueba. Los resultados de este experimento de fusión a bajo nivel se asemejan a los obtenidos por el sistema monomodal de reconocimiento de emociones en voz con tres capas convolucionales.

a) MLLF_3CC_16_13 b) MLLF_3CC_32_18

c) MLLF_3CC_64_15

Figura 38. Fusión en bajo nivel. BD3 y RAVDESS

Fusión en EMO-MX-NAO

La Figura 39 ilustra el resultado de la unión del espectrograma-expresión facial en una imagen de 384x192 pixeles.

Figura 39. Fusión a bajo nivel EMO-MX-NAO

Puesto que la base de datos de expresiones faciales es mayor que la de audio, fueron descartadas aquellas imágenes sin par, teniendo la distribución de la Tabla 15.

Tabla 15. Distribución de imágenes para la base de datos EMO-MX-NAO

Emoción	Imágenes de entrenamiento	Imágenes de validación	Imágenes de prueba
Enojo	68	12	11
Aversión	68	12	11
Miedo	68	12	11
Alegría	68	12	11
Tristeza	68	12	11
Sorpresa	68	12	11

La Tabla 16 muestra los resultados de diferentes entrenamientos para el sistema multimodal de reconocimiento de emociones con fusión de bajo nivel aplicado al conjunto de imágenes EMO-MX-NAO.

Tabla 16. Modelos de fusión de bajo nivel EMO-MX-NAO

Nombre del modelo	*Epochs*	*Batch_size*	*Training accuracy*	*Validation accuracy*	*Test accuracy*
MLLF_NAO_3CC_16_15	15	16	0.96	0.75	0.72
MLLF_NAO_3CC_32_15	15	32	0.95	0.79	0.71
MLLF_NAO_3CC_64_13	13	64	0.87	0.73	0.72

Las matrices de confusión de los modelos descritos en la Tabla 16, Figura 40, muestran un mayor balance en la predicción para todas las clases respecto a los resultados obtenidos con el *dataset* DB3 y RAVDESS. Se selecciona el modelo MLLF_NAO_3CC_64_13 como el de mejor desempeño para este experimento.

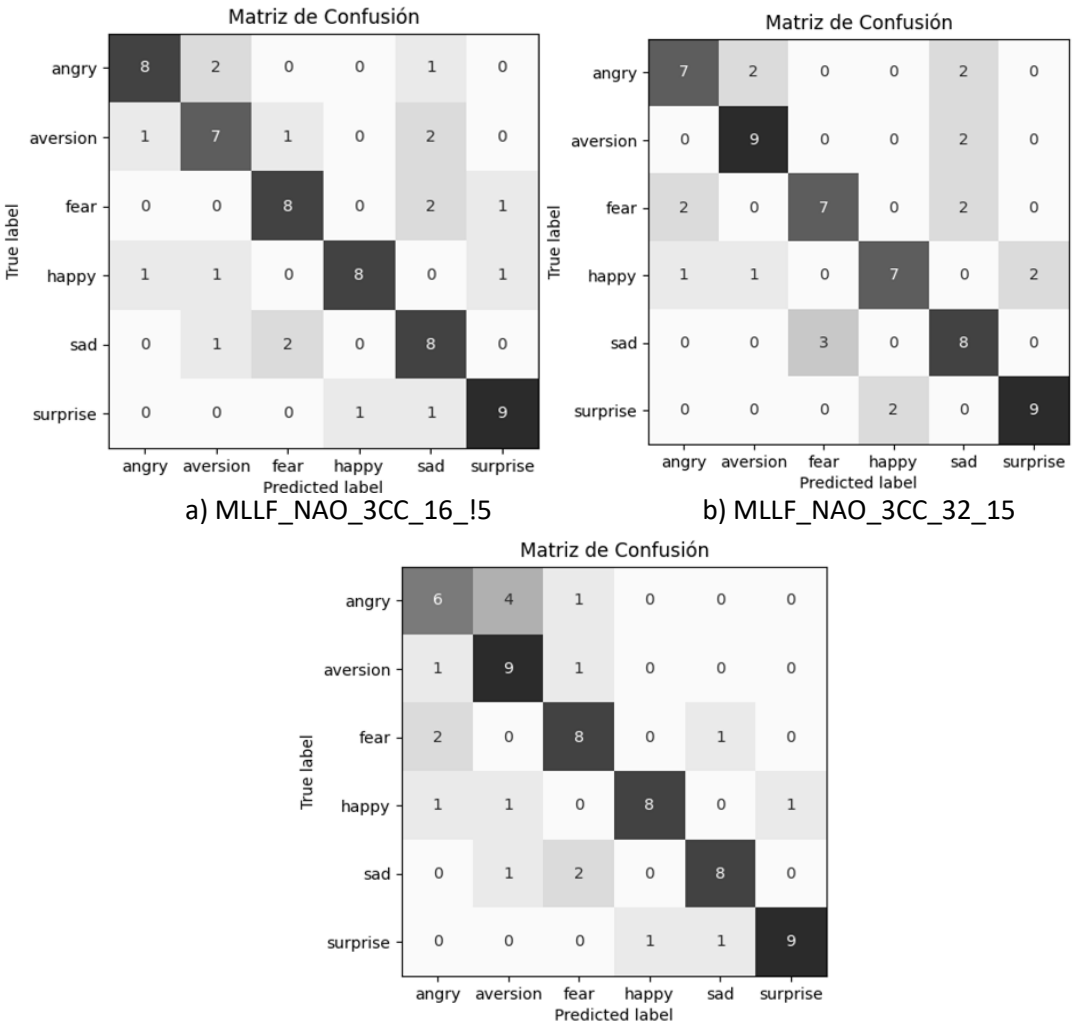


Figura 40. Matrices de confusión de los modelos de fusión de bajo nivel en EMO-MX-NAO

4.3.2 Fusión en nivel medio

La Figura 41 ilustra la arquitectura utilizada para la fusión de nivel medio en el sistema multimodal de reconocimiento de emociones. Para cada rama se define una sub-red de tres capas convolucionales y una capa densa cuyas salidas se concatenan y se convierten en la entrada a una tercera capa densa para finalmente pasar a la capa de clasificación.

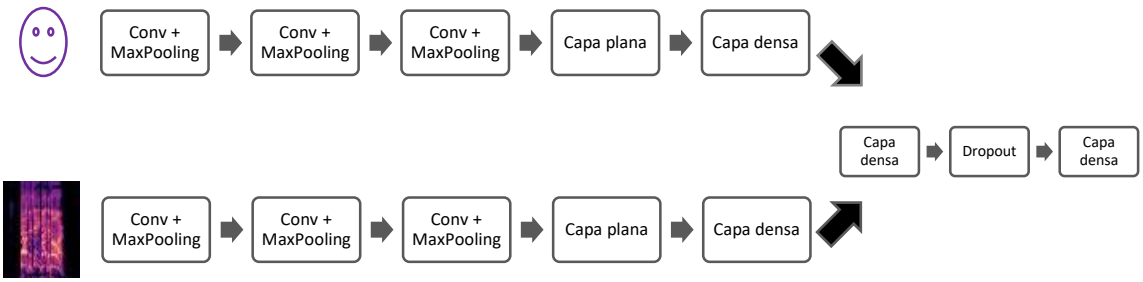


Figura 41. Arquitectura del sistema multimodal de reconocimiento de emociones con fusión media

Las entradas de cada rama corresponden a las imágenes de las expresiones faciales y los espectrogramas utilizados en fusión de bajo nivel, pero sin combinar para poder ingresar correctamente a cada sub-red.

Fusión en BD3 y RAVDESS

La Tabla 17 contiene los resultados de los entrenamientos para el sistema multimodal de reconocimiento de emociones con fusión de nivel medio aplicado al conjunto de datos BD3 y RAVDESS.

Tabla 17. Modelos de fusión de nivel medio para BD3 y RAVDESS

Nombre del modelo	*Epochs*	*Batch_size*	*Training accuracy*	*Validation accuracy*	*Test accuracy*
MMLF_3CC_16_13	13	16	0.70	0.59	0.54
MMLF_3CC_32_18	18	32	0.74	0.55	0.60
MMLF_3CC_64_15	15	64	0.78	0.58	0.60

Los modelos de fusión en nivel medio presentan un desempeño menor a los de fusión en bajo nivel y mayor dispersión en la clasificación, tal como lo indican las matrices de confusión de los modelos descritos en la Tabla 17, Figura 42, para las bases de datos BD3 y RAVDESS.

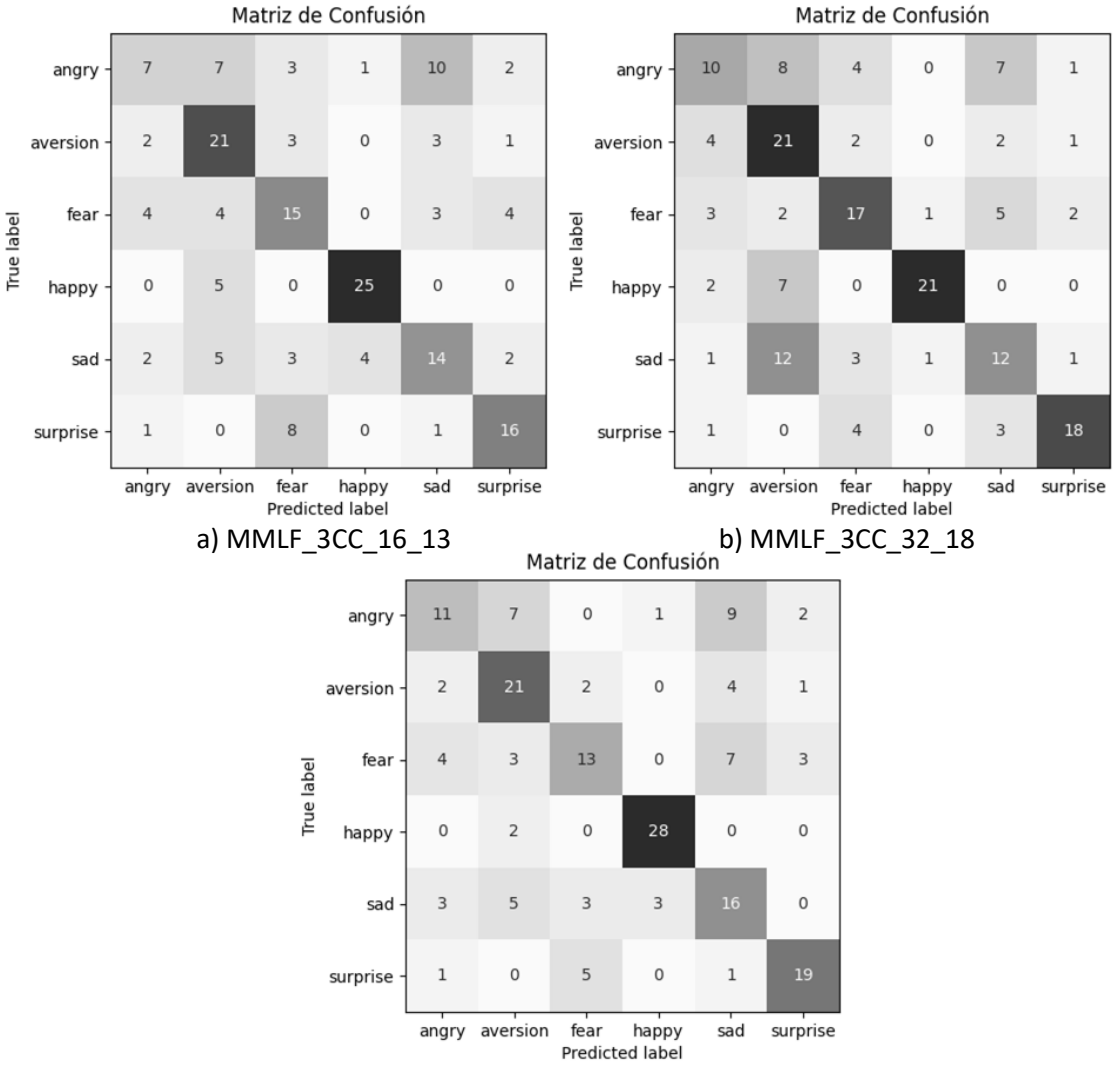


Figura 42. Matrices de confusión de los modelos de fusión a nivel medio para BD3 y RAVDESS

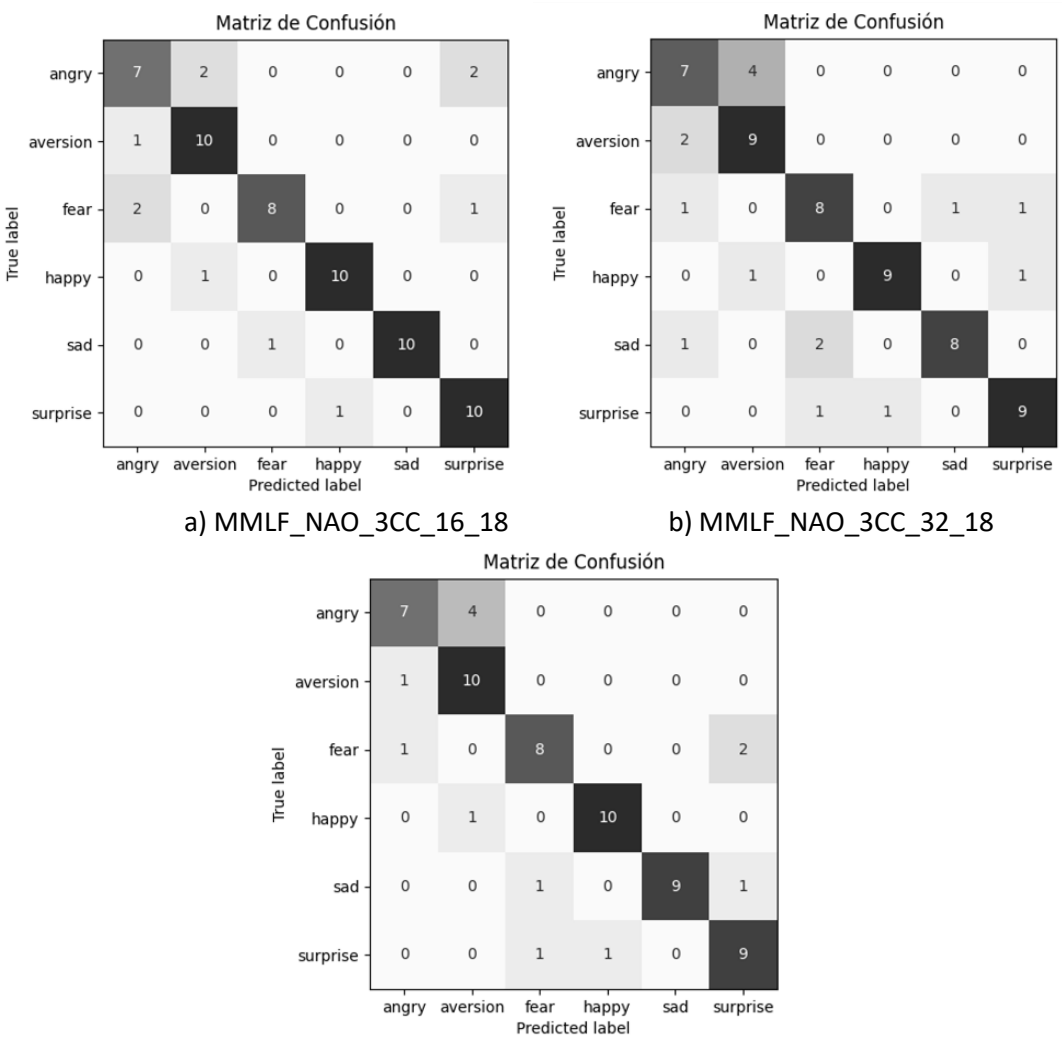
Fusión en EMO-MX-NAO

La Tabla 18 muestra los resultados obtenidos al efectuar diferentes entrenamientos para el sistema multimodal de reconocimiento de emociones con fusión de nivel medio para la base de datos EMO-MX-NAO

Tabla 18. Modelos de fusión de nivel medio para EMO-MX-NAO

Nombre del modelo	*Epochs*	*Batch_size*	*Training accuracy*	*Validation accuracy*	*Test accuracy*
MMLF_NAO_3CC_16_18	18	16	0.99	0.82	0.83
MMLF_NAO_3CC_32_18	18	32	0.99	0.76	0.75
MMLF_NAO_3CC_64_18	18	64	0.99	0.79	0.80

La Figura 43 muestra las matrices de confusión de los modelos descritos en la Tabla 18. Se observa un mayor balance en la clasificación de las emociones respecto a los modelos de mediano nivel para BD3 y RAVDESS



a) MMLF_NAO_3CC_16_18

b) MMLF_NAO_3CC_32_18

c) MMLF_NAO_3CC_64_18

Figura 43. Matrices de confusión para modelos fusión nivel medio en EMO-MX-NAO

4.3.3 Fusión en alto nivel

La Figura 44 ilustra la arquitectura seleccionada para el sistema multimodal de reconocimiento de emociones en fusión de alto nivel. Las constantes α y β representan los porcentajes asignados a cada rama para la clasificación de emociones. Se han seleccionado los modelos con mejor desempeño monomodal en clasificación de emociones en voz y expresiones faciales a la entrada de cada fusión.

Figura 44. Arquitectura de fusión de alto nivel

Se utilizó el mismo conjunto de datos de prueba empleado en la fusión de nivel medio para cada experimento.

Fusión en BD3 y RAVDESS

Para el sistema multimodal de reconocimiento de emociones con fusión en nivel alto se ha seleccionado los modelos CNN_EF_5CC_256_2000 y CNN6_Voice_RAV_64_25 previamente entrenados en la clasificación de expresiones faciales y emociones en voz, respectivamente.

La Tabla 19 contiene los diferentes valores de precisión en imágenes de prueba bajo diferentes valores de α y β. Se observa que los mejores desempeños se presentan con valores de α en un intervalo de (0.50, 0.60)

Tabla 19. Fusión de alto nivel para BD3 y RAVDESS

α	β	*Test accuracy*
0.05	0.95	0.66
0.10	0.90	0.67
0.15	0.85	0.67
0.20	0.80	0.66
0.25	0.75	0.68
0.30	0.70	0.71
0.35	0.65	0.72
0.40	0.60	0.74
0.45	0.55	0.79
0.50	0.50	0.81
0.55	0.45	0.81
0.60	0.40	0.81
0.65	0.35	0.80
0.70	0.35	0.78
0.75	0.25	0.78
0.80	0.20	0.76
0.85	0.15	0.76
0.90	0.10	0.75
0.95	0.05	0.73

Para el conjunto de datos BD3 y RAVDESS el mejor porcentaje de precisión se alcanza con el sistema multimodal de reconocimiento de emociones con fusión de alto nivel.

Fusión en EMO-MX-NAO

Para el sistema multimodal de reconocimiento de emociones con fusión en nivel alto, bajo la base de datos EMO-MX-NAO se ha seleccionado los modelos

CNN_EF_NAO_3CC_64_120 y CNN6_Voice_NAO_16_40 para la clasificación de expresiones faciales y emocione en voz, respectivamente.

La Tabla 20 contiene los valores de precisión obtenidos para los datos de prueba bajo diferentes valores de α y β.

Tabla 20. Fusión de alto nivel para EMO-MX-NAO

α	β	*Test accuracy*
0.10	0.90	0.58
0.15	0.85	0.59
0.20	0.80	0.61
0.25	0.75	0.64
0.30	0.70	0.70
0.35	0.65	0.71
0.40	0.60	0.80
0.45	0.55	0.83
0.50	0.50	0.88
0.55	0.45	0.91
0.60	0.40	0.91
0.65	0.35	0.90
0.70	0.35	0.91
0.75	0.25	0.88
0.80	0.20	0.88
0.85	0.15	0.86
0.90	0.10	0.86
0.95	0.05	0.85

Se observa mejor desempeño de los modelos con valores de α en un intervalo de (0.50, 0.65) De la misma forma, el sistema multimodal de reconocimiento de emociones con fusión en alto nivel muestra mejores resultados para la base de datos EMO-MX-NAO

en comparación con los resultados obtenidos con las bases de datos preexistentes BD3 y RAVDESS.

Todos los códigos utilizados durante la investigación se encuentran disponibles en el siguiente repositorio de GitHub para su revisión:

https://github.com/CristellTL/CNN-Multimodal-Reconocimiento-de-Emociones

Capítulo 5

ANÁLISIS DE RESULTADOS Y CONCLUSIONES

Desde su aparición, la robótica social ha tenido un alto impacto en la vida de los seres humanos. Los robots están siendo dotados cada vez más de habilidades que habrían podido pensarse exclusivas del hombre, tal como comunicarse y tomar decisiones con base en experiencias. De esto último surge la importancia de la Inteligencia Artificial aplicada a la robótica social.

El reconocimiento de emociones en un enfoque de Interacción Humano Robot supone un reto que permite una comunicación más efectiva, afectiva y natural puesto que los robots podrán comunicarse basándose en el estado de ánimo de las personas.

El desarrollo de esta investigación permitió demostrar que las emociones pueden clasificarse con porcentajes de precisión más altos, proporcionales al número de señales analizadas para la clasificación y la calidad de las muestras contenidas en la base de datos.

La Tabla 21, contiene un resumen de los modelos monomodales para la clasificación de expresiones faciales y emociones en voz con mejores precisiones en el conjunto de prueba tanto para bases de datos preexistentes como para la base de datos generada durante la investigación.

Tabla 21. Comparación de los modelos monomodales de mejor desempeño revisados

Nombre del modelo	*Base de datos utilizada*	*Señal analizada*	*Precisión en conjunto de prueba*
CNN_EF_3CC_256_2000	BD3	Expresión facial	0.66
CNN_EF_NAO_3CC_64_120	EMO-MX-EF-NAO	Expresión facial	0.87

CNN6_Voice_RAV_64_25	RAVDESS	Emociones en habla	0.63
CNN3_Voice_RAV_32_25	RAVDESS	Emociones en habla	0.61
CNN3_Voice_NAO_32_45	EMO-MX-SP-NAO	Emociones en habla	0.48
CNN6_Voice_NAO_16_40	EMO-MX-SP-NAO	Emociones en habla	0.56

Las expresiones faciales mostraron un porcentaje de clasificación superior al utilizar un modelo entrenado con la base de datos generada a partir de la interacción con el robot humanoide NAO, lo que es atribuible al entorno controlado donde se tomaron las fotografías, el ambiente de confianza que generó el robot y la dimensionalidad de las imágenes en comparación con las bases de datos preexistentes.

En contraste, los modelos de clasificación de emociones en voz muestran mejores resultados con la base de datos RAVDESS en comparación con la base de datos EMO-MX-SP-NAO, sin embargo, no puede realizarse una comparación válida que indique los parámetros que influyen en la precisión obtenida con la base de datos generada debido a los idiomas diferentes bajo los que fueron creados los *datasets.*

De forma general, la clasificación de emociones muestra mejores resultados cuando se realiza sobre las expresiones faciales en comparación con los análisis realizados para las emociones en habla.

La Tabla 22, muestra el resumen de los modelos multimodales para la clasificación de emociones empleando las dos señales: expresión facial y voz, bajo bases de datos preexistentes y la base de datos EMO-MX-NAO.

Se observa que el desempeño de los modelos bajo los tres niveles de fusión es más alto al ser entrenado y evaluado con la base de datos EMO-MX-NAO, lo cual indica una alta calidad en el *dataset* generado.

El nivel de fusión que presenta mejores resultados en precisión en el conjunto de prueba es el nivel alto, donde se encuentra la precisión más alta obtenida durante toda la investigación: 90%.

Tabla 22. Comparación de los modelos multimodales analizados

Nombre del modelo	*Base de datos utilizada*	*Tipo de fusión*	*Precisión en conjunto de prueba*
MLLF_3CC_16_30	BD3 y RAVDESS	Bajo nivel	0.66
MLLF_NAO_3CC_64_13	EMO-MX-NAO	Bajo nivel	0.72
MMLF_3CC_64_30	BD3 y RAVDESS	Nivel medio	0.60
MMLF_NAO_3CC_16_18	EMO-MX-NAO	Nivel medio	0.83
HightLevelFusion	BD3 y RAVDESS	Alto nivel	0.81
HightLevelFusion-NAO	EMO-MX-SP-NAO	Alto nivel	0.91

Con los datos analizados en las tablas anteriores, se demuestra que un sistema de reconocimiento de emociones bajo arquitecturas de redes neuronales convolucionales con alto porcentaje de precisión, se obtienen al efectuar análisis multimodales con más de una señal de entrada

Como trabajo a futuro, se plantea la integración de una tercera señal de entrada al sistema multimodal, siendo esta un conjunto de muestras históricas de la frecuencia cardiaca presente en la persona durante la interpretación de la emoción.

Referencias

- Ahmed J. Obaid, Hassanain K. Alrammahi. (2023). An Intelligent Facial Expression Recognition System Using a Hybrid Deep Convolutional Neural Network for Multimedia Applications. *Applied Sciences*, *13*(21), 12049. DOI: 10.3390/app132112049
- Akram, S., Alhajlah, M., & Mahmood, A. (2023). Hybrid Facial Emotion Recognition Using CNN-Based Features. *Applied Sciences*, *13*(9), 5572.DOI: 10.3390/app13095572
- Asociación Española contra el Cáncer. (2010). *Las emociones, comprenderlas para vivir mejor. Guías y protocolos*. Madrid, España.
- Breazeal, C. (2004). *Designing Sociable Robots*. MIT Press.
- Chóliz Montañés, M. (2005). *Psicología de la emoción: El proceso emocional*. Departamento de Psicología Básica, Universidad de Valencia.
- Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357-366. DOI: 10.1109/TASSP.1980.1163420
- Ekman, P., & Friesen, W. V. (1971). *Constants across cultures in the face and emotion*. Journal of Personality and Social Psychology, 17(2), 124-129.DOI: 10.1037/h0030377
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., ... & Maisonnier, B. (2009). Mechatronic design of NAO humanoid. *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*, 769-774.DOI: 10.1109/ROBOT.2009.5152516
- Han, J., Zhang, D., Cheng, G., Guo, L., & Ren, J. (2015). Object detection in optical remote sensing images based on weakly supervised learning and high-level

feature learning. *IEEE Transactions on Geoscience and Remote Sensing, 53*(6), 3325-3337.

- Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech* (pp. 223-227).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. En *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. En *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kanade, T., Cohn, J. F., & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition* (FG'00), 46-53. DOI: 10.1109/AFGR.2000.840611
- Kim, S., & Lee, S.-P. (2023). *A BiLSTM–Transformer and 2D CNN Architecture for Emotion Recognition from Speech*. Electronics, 12, 4034. DOI: 10.3390/electronics12194034
- Kudoh, S., Komoriya, K., Inaba, M., & Inoue, H. (2005). Whole-body humanoid robot control through ZMP manipulation. *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, 1316-1323. DOI: 10.1109/ROBOT.2005.1570300
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. Nature, 521(7553), 436-444. DOI: 10.1038/nature14539
- Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3), 1191-1214. DOI: 10.1109/TAFFC.2020.2981446
- Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial

and vocal expressions in North American English. *PloS one*, 13(5), e0196391. DOI: 10.1371/journal.pone.0196391

- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 94-101. DOI: 10.1109/CVPRW.2010.5543262
- Michaud, F., Boissy, P., Labonte, D., Corriveau, H., Grant, A., Lauria, M., ... & Cloutier, R. (2007). Exploratory design and evaluation of a homecare robot for older adults. *Autonomous Robots, 22*(4), 401-417. DOI: 10.1007/s10514-007-9055-8
- Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2), 98-100.
- Moro, M., Obaid, M., & Young, S. (2012). NAO Robot as a Teaching Assistant for University Courses. *Proceedings of the 24th Australian Computer-Human Interaction Conference*, 448-451. DOI: 10.1145/2414536.2414616
- Murugan, H. (2020). Speech Emotion Recognition Using CNN. *International Journal of Psychosocial Rehabilitation*, 24. DOI: 10.37200/IJPR/V24I8/PR280260
- Pan, J., Fang, W., Zhang, Z., Chen, B., Zhang, Z., & Wang, S. (2023). Multimodal Emotion Recognition based on Facial Expressions, Speech, and EEG. *IEEE Open Journal of Engineering in Medicine and Biology*. Advance online publication. DOI: 10.1109/OJEMB.2023.3240280
- Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, 98-125. DOI: 10.1016/j.inffus.2017.02.003
- Rahaman, Muhammad. (2024). *A 3D CNN Model with Multi-Feature Fusion for Enhancing Human Emotion Recognition from Speech.*
- Sánchez Martín, F., Jiménez Schlegl, P., Millán Rodríguez, F., Salvador Bayarri, J., Monllau Font, V., Palou Redorta, J., & Villavicencio Mavrich, H. (2007,

marzo). Historia de la robótica: de Arquitas de Tarento al Robot da Vinci. *Actas Urológicas Españolas.*

- Saranya Rajan, P., Chenniappan, P., & Devaraj, S. (2020). Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM. *IET Image Processing, 14*(10), 2009-2016. DOI: 10.1049/iet-ipr.2019.1188
- Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2010). The INTERSPEECH 2010 paralinguistic challenge. In *Eleventh Annual Conference of the International Speech Communication Association.*
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*
- SoftBank Robotics. (n.d.). NAO the Humanoid and Programmable Robot. SoftBank Robotics. Retrieved from https://www.softbankrobotics.com/emea/en/nao
- Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5200-5204). IEEE. DOI: 10.1109/ICASSP.2016.7472669
- Wang, X., & Gupta, A. (2018). Videos as space-time region graphs. En *European Conference on Computer Vision (ECCV)*.
- Yu, Z., & Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 435-442. DOI: 10.1145/2818346.2830595
- Zhao, Z., Zou, Y., & Zhang, J. (2023). *Speech emotion recognition using convolutional neural networks and multi-head convolutional transformer*. Sensors, 23(13), 6212. DOI: 10.3390/s23136212