ESCUELA DE INGENIERÍA DE FUENLABRADA

GRADO EN INGENIERÍA EN SISTEMAS AUDIOVISUALES Y MULTIMEDIA

TRABAJO FIN DE GRADO

Identificación de emociones en señales de voz a partir de técnicas de aprendizaje automático

Autor: María Martínez Ruiz

Tutor: David Gualda Gómez

Curso académico 2023/2024

AGRADECIMIENTOS

Parecía que nunca iba a llegar este momento, pero si algo me ha enseñado la carrera es que todo esfuerzo tiene su recompensa y con esta memoria por fin, puedo poner un punto final a esta larga pero bonita etapa como estudiante de ISAM.

En primer lugar, mi más sincero agradecimiento a mi tutor, David, por acompañarme y no desesperar durante estos casi 2 años de TFG, ha sido un placer trabajar y aprender contigo.

Gracias también a mi familia por ser los que más habéis sufrido las consecuencias de mis momentos menos buenos, es una verdadera suerte crecer en un entorno lleno de apoyo y confianza, gracias en especial a mi hermano Adrián, por ser mi referente desde bien pequeña y por ir abriéndome puertas que jamás me hubiese atrevido a cruzar.

Gracias a la carrera, por demostrarme que no es imposible sacar un 10, pero que tampoco lo es sacar un 0… Recordaré con cariño cada risa y cada lagrima vivida en clase, pero gracias sobre todo por haberme ofrecido la oportunidad de irme un año a estudiar fuera. Nunca me hubiese imaginado que se podía aprender tanto en tan poco tiempo ni que iba a conocer gente que se alegra más por mí que yo misma, Lucía, Dani y resto de grupo Erasmus, me lo habéis hecho muy cuesta abajo, todo lo bueno que pueda decir de nuestra experiencia, se queda corto.

Y, por último, gracias a todos mis compañeros de clase, Lucía, Thalía, Emilio y especialmente Carlos, como hemos sufrido y como hemos disfrutado juntos. Gracias chicos, por medio enseñarme a jugar al Volley, por hacerme más amenas las tardes en la biblioteca y por formar equipo desde el día 1 de universidad. No me olvido de ti Adri, que tu supiste unirnos y hacernos reír más y mejor que nadie, repetiría esta aventura mil veces más solo por volver a conocerte.

A todos vosotros y a los que olvido nombrar, gracias de corazón.

RESUMEN

En este Trabajo Fin de Grado (TFG), se detallan todos los aspectos necesarios para el estudio y desarrollo de un modelo de aprendizaje destinado a clasificar estados emocionales a partir de señales de voz. El proceso implica la exploración y la implementación de diferentes redes neuronales *feedforward* para dicha tarea de reconocimiento emocional. Se llevan a cabo al menos treinta bancos de pruebas utilizando tres bases de datos públicas para determinar la idoneidad de cada red neuronal.

El procedimiento sigue varias etapas, que incluyen, la selección del conjunto de datos, su tratamiento previo y posterior, la identificación de características relevantes y, finalmente, la elección de un algoritmo de aprendizaje que optimice al máximo la precisión en la estimación emocional de las señales de audio.

Para evaluar la eficacia de los modelos, se utilizan diferentes métricas de evaluación y los resultados obtenidos son interpretados y comparados con el objetivo de identificar la red neuronal que proporciona la mayor precisión en la clasificación.

Además, se exploran posibles aplicaciones prácticas de este modelo en otros contextos. Este enfoque amplio pero detallado, permite una comprensión del impacto y la aplicabilidad del modelo propuesto.

ABSTRACT

In this Final Degree Project (FDP), all the necessary aspects for the study and development of a learning model to classify emotional states from voice signals are detailed. The process involves the exploration and implementation of different feedforward neural networks for such emotional recognition task. At least thirty test beds are conducted using three public databases to determine the suitability of each neural network.

The procedure follows several stages, which include, the selection of the dataset, its pre and post processing, the identification of relevant features and, finally, the choice of a learning algorithm that optimizes to the maximum the accuracy in the emotional estimation of the audio signals.

To evaluate the effectiveness of the models, different evaluation metrics are used, and the results obtained are interpreted and compared with the aim of identifying the neural network that provides the highest classification accuracy.

In addition, possible practical applications of this model in other contexts are explored. This broad but detailed approach allows an understanding of the impact and applicability of the proposed model.

Índice de Contenidos

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

ACRÓNIMOS

ANN	Artificial neural network
AUC-ROC	Área bajo la curva - receiver Operating Characteristic
BPTT	*Backpropagation through time*
CNN	Convolutional neural networks
CREMAD	*Crowd-sourced Emotional Mutimodal Actors Dataset*
DB	Decibel
FFT	Fast Fourier transform
Fs	Sampling frequency
FT	Fourier transform
GB	Gigabyte
GMM	Modelos Gaussianos Mixtos
Gpu	Graphics processing unit
HMM	Hidden Markov Models
IA	Artificial Intelligence
KNN	K-nearest neighbors
LFPC	Log Frequency speech power Coefficients
LSTM	Long short-term memory
MATLAB	Matrix laboratory
MFCC	Mel Frequency Cepstral Coefficients
ML	Machine learning
MPL	Multiprotocol Label Switching
NN	Neural network
RAVDESS	*Ryerson Audiovisual Database of Emotional Speech and Song*)
RMS	*Root Mean Square*
RNA	Artificial Neural Network
RNN	recurrent neural networks
SAVEE	*Surrey Audiovisual Expressed Emotion*
SHAP	SHapley Additive exPlanations
STFT	Short Time Fourier Transform
SVM	Support vector machine
ZCR	zero-crossing rate

1. INTRODUCCIÓN

La detención de emociones no es fácil para una maquina porque tampoco lo es para nosotros. Leí sobre un caso de agresión sexual hacia una joven que aparentemente no reflejaba nerviosismo, ansiedad ni otros indicios que indicasen que había estado en peligro, al parecer desde niña había vivido en otros países donde son menos expresivos, por eso se percibía su historia como fría y distante. La realidad era que esa chica tenía todos esos síntomas, pero no eran visibles.

Las emociones dependen de la persona, también de su edad o idioma entre otros. Hay tantos factores que pueden determinar una emoción, que se hace muy complejo llevar la investigación del reconocimiento de emociones a la práctica.

El ser humano es un ser emocional, utilizamos cuerpo y voz como vehículo de expresión, sin embargo, muchas veces no son necesarios ambos para entender dicho estado. Por ejemplo, no se requiere oír la voz, pero es de gran ayuda a la hora de saber interpretar una emoción.

Dada la dificultad, algunos autores también diferencian las emociones complejas, que surgen como una combinación de emociones básicas. Robert Plutchik [1], considera 8 como base y estas las utiliza para desarrollar su modelo llamado la Rueda de las Emociones, un mapa que permite ver las diferentes combinaciones de estas.

Existen muchas emociones y muchas formas de definirlas y representarlas, establecer una clasificación de estas no es una tarea fácil, pues no todos los autores hablan del mismo tipo y número de emociones. Piaget [1], por ejemplo, nombra más de 400 diferentes, Goleman habla de la ira, habla del amor, la vergüenza... En 1994 fue Paul Ekman el que identificó un conjunto de 6 emociones básicas: Miedo (*fear*), tristeza (*sadness*), alegría (*joy*), ira (*anger*), sorpresa (*surprise*) y asco (*disgust*). Posteriormente, Ekman añadiría una séptima, el desprecio (*contempt*). Se las considera básicas por estar ligadas a la supervivencia de los individuos y por su universalidad y su presencia en distintas culturas.

Aunque muchos otros psicólogos consideran también otras emociones, como la neutralidad (*neutral*), la mayoría coincide con que las básicas no superan las 10.

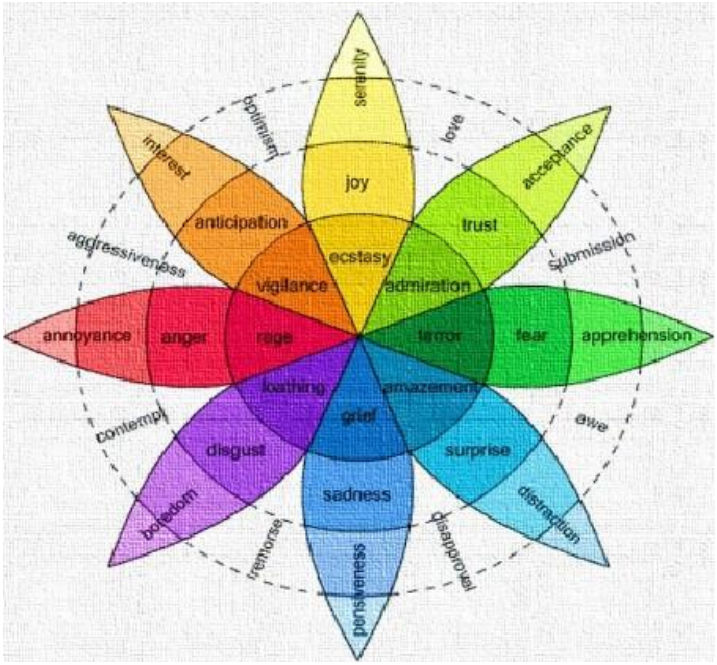


Figura 1. Rueda de las emociones de Plutchik (1980) [2].

En esta rueda, Plutchik describe la relación entre las distintas emociones representadas con colores distintos. En el interior de la estrella, los colores más intensos equivalen a las emociones primarias, las 8 mencionadas anteriormente [2]. Los colores difuminados, más alejados del centro en función de su intensidad, definen las emociones secundarias (cuanto más alejadas, más secundarias), y las que ocupan los huecos blancos son el resultado de mezclar dos primarios.

1.2. Objetivos y enfoque

El objetivo principal del trabajo es la elaboración de un sistema de reconocimiento automático de emociones a partir de señales voz.

"Los programadores deben considerar la perspectiva del afecto al crear un software destinado a interactuar con humanos", escribió Rosalind Picard, profesora del instituto tecnológico de Massachusetts.

En 1995, Rosalind introdujo la teoría de dotar a una máquina de respuesta emocional para conseguir que esta sea verdaderamente inteligente [3][4].

Cada vez son más las horas que pasamos con nuestro ordenador, Tablet, smartphone... y ante ellos no fingimos, por eso sería interesante que estos dispositivos fueran capaces de averiguar nuestro estado de ánimo y conocer cómo nos sentimos para actuar en consecuencia [3].

¿A dónde se quiere llegar con el análisis de emociones mediante la voz?

Yo, como persona, me puedo imaginar lo que significa una pausa dentro de una conversación, sé interpretar un suspiro, risa o un simple "um,"uf", sin embargo, mi smartphone se mantendría frío ante cualquiera de estos sonidos. Si una máquina fuese capaz de analizar esos pequeños detalles del habla se podría por ejemplo mejorar la experiencia del usuario al usar el reconocimiento de voz del asistente de un dispositivo o de Alexa.

Con la intención de dar solución a esto, se desarrollará una red neuronal en Matlab, diseñada para clasificar automáticamente las emociones de una persona en su discurso. Para el desarrollo se utilizará una base de datos de emociones básicas que se describirán en un apartado posteriormente.

La finalidad de este trabajo es, por tanto, adquirir todos los conocimientos necesarios para diseñar una máquina que será entrenada, automatizada y dotada con la capacidad de procesar y clasificar con precisión unas muestras de audio con emociones, que serán evaluadas con el objetivo de determinar sus prestaciones y limitaciones.

En función de esto, la máquina conseguiría una respuesta similar a la que tendría un ser humano, mejorando así la comunicación y haciéndola lo más parecida posible a la interacción entre personas.

1.3. Organización de la memoria

El proyecto se ha estructurado en 5 apartados para facilitar la lectura y comprensión de la memoria, a continuación, se presenta un breve resumen de cada capítulo:

Capítulo 1:

En este primer capítulo, se establecen los objetivos específicos del proyecto. Se describen las herramientas y tecnologías utilizadas en la investigación, proporcionando una visión general del marco de trabajo.

Capítulo 2:

Con un pequeño resumen del estado del arte en la predicción de emociones, se exploran las características del habla, comprendiendo el sistema fonador y los elementos más relevantes del proceso del habla, profundizando en su relación con la expresión emocional. Además, se examinan las características generales de las señales de voz proporcionando una base sólida para la comprensión de las señales de voz utilizadas en el reconocimiento.

Capítulo 3:

Este apartado se dedica a un estudio más teórico de la inteligencia artificial abordando sus inicios, del *machine learning*, *Deep learning* y las redes neuronales en general. Se presenta el concepto de computación afectiva, destacando su importancia en el reconocimiento de emociones. También se enfatiza la relevancia de las bases de datos y se exploran métricas de evaluación fundamentales para la evaluación de sistemas de clasificación como el que se desarrolla en este proyecto.

Capítulo 4:

En este capítulo, se detalla la base de datos utilizada, así como el proceso de recolección para construir el conjunto de datos deseado. Se explica la división del conjunto para alimentar el modelo, junto con el preprocesado de las señales de voz la selección de características relevantes en el desarrollo.

Capítulo 5:

Este apartado se centra en los detalles de la implementación del proyecto, desde la configuración del modelo hasta la obtención de resultados. Se presentan y analizan los resultados obtenidos, proporcionando una visión integral del rendimiento del sistema de reconocimiento de emociones en señales de voz.

Capítulo 6:

En este capítulo final, se presentan las conclusiones derivadas de la investigación. Se sintetizan las capacidades del modelo para identificar emociones y se plantean nuevas líneas de investigación en el campo de la computación afectiva a partir de señales de voz.

1.4. Diagrama de Gantt

MESES	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Mes 7	Mes 8	Mes 9	Mes 10	Mes 11	Mes 12
INVESTIGACIÓN												
DESARROLLO MATLAB												
ESTUDIO RESULTADOS												
DOCUMENTACIÓN												

Tabla 1. Diagrama de Gantt

1.5. Herramientas utilizadas

La implementación y evaluación del proyecto se ha llevado a cabo en MATLAB R2023b, entorno versátil y eficiente que ha demostrado ser fundamental para el desarrollo de la red neuronal y la manipulación de señales de audio. Se eligió este entorno por la variedad de *Toolboxes* especializadas que ofrece tanto para el procesamiento de señales como para el entrenamiento de redes neuronales; por las herramientas gráficas que incluye para facilitar la visualización y análisis de los datos y, porque en una balanza Matlab-Python, se consideró que MATLAB puede optimizar el proceso y resultar en una ejecución

más eficiente y sencilla de algoritmos complejos, especialmente si se implican volúmenes grandes de datos.

La *Toolbox* de Procesamiento de Señales ha sido esencial para facilitar operaciones críticas como el filtrado, preénfasis y segmentación temporal, preparando así las señales para su posterior análisis. La *Toolbox* de Audio también ha sido clave para la manipulación de las señales de audio, utilizando funciones especializadas, como '*cromagram*' y '*spectrogram*' que contribuyeron a la extracción de características relevantes para el análisis emocional.

La construcción del modelo de reconocimiento se llevó a cabo con la *Toolbox* de Aprendizaje Profundo, utilizando la función '*patternet*'. Esta herramienta incorpora las funciones necesarias para llevar a cabo los procesos de validación y evaluación.

MATLAB también ofrece otras herramientas gráficas que se han aprovechado para visualizar y analizar de manera efectiva los resultados generados en el desarrollo del proyecto. Estas capacidades gráficas fueron cruciales para una comprensión más profunda y una presentación más clara de los resultados obtenidos.

2. ESTADO DEL ARTE

Antes de iniciar este estado del arte sobre el proceso de reconocimiento de emociones en la voz, es esencial familiarizarse con algunos conceptos básicos sobre los aspectos psicológicos, biológicos y lingüísticos de las emociones.

2.1. El aparato fonador humano

El sistema fonador se refiere a la colección de órganos en el cuerpo humano responsables de producir el sonido emitido durante el habla. Los órganos que participan en este proceso cumplen otras funciones claves para la supervivencia del ser humano, por ejemplo, la función principal de la laringe es evitar que los alimentos y líquidos entren a la tráquea, pero también juega un papel esencial en el proceso de la producción del habla humana.

El sistema vocal humano se divide en 3 grupos de órganos: los órganos de respiración, los de fonación y los de articulación. Estos, contienen los tres elementos fundamentales que establecen las leyes de la acústica: un cuerpo vibrante, un medio elástico que propague las vibraciones y una caja de resonancia que las amplifique para que puedan ser percibidas por el oído, a continuación, se ve más detalladamente [5]:

El primer grupo tiene la función de proporcionar oxígeno al cuerpo y también de producir sonidos a partir del aire expulsado. Forman parte de este grupo la faringe, laringe, tráquea, pulmones y diafragma [6].

Con los pulmones ya llenos de aire, entran en juego los órganos de fonación, que incluyen la laringe, las cuerdas vocales, la faringe, la cavidad nasal y la cavidad bucal (estos 2 últimos, conocidos como cavidades resonantes), utilizan el aire para generar vibraciones que el sistema auditivo interpretará como sonidos [6].

Por último, el tercer grupo, que comprende la glotis, el paladar, la lengua, los dientes y los labios, toma el sonido ya generado, lo amplifica y modula,

otorgándole los matices necesarios para que la voz se traduzca en un significado a través de las palabras.

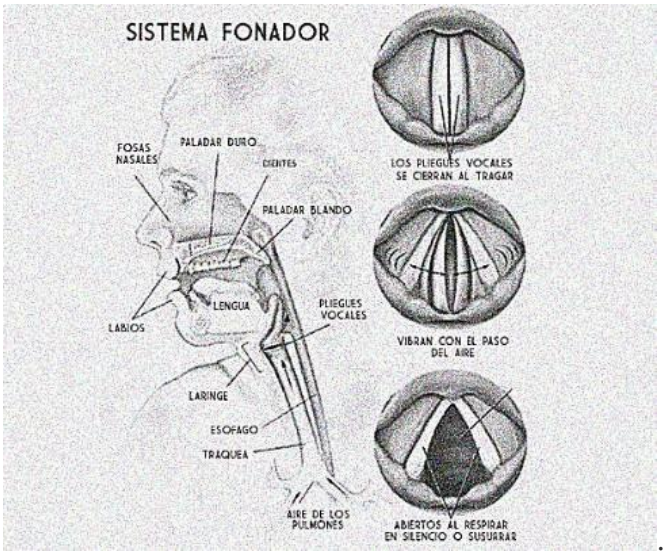


Figura 2. Sistema fonador humano [7].

La voz es por lo tanto el resultado de un proceso realizado por el aparato fonador. Se trata de una señal acústica que se forma con el movimiento de moléculas de aire. Esta onda se representa como una onda sinusoidal, como la que se muestra en la figura 3.

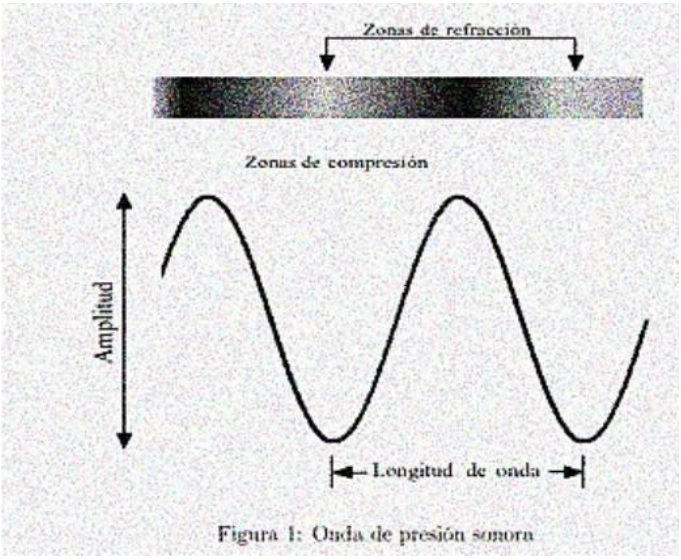


Figura 3. Onda de presión sonora [7].

2.2. El reflejo de las emociones en la voz

La voz de cada persona es única, la onda sonora que se genera con la vibración de las cuerdas vocales y que atraviesa los órganos que se han visto en el apartado anterior son diferentes para cada individuo. Además, hay otros factores que influyen de forma directa en la voz.

En 1967, Albert Mehrabian llevó a cabo un estudio con el objetivo de cuantificar la influencia de las palabras, el lenguaje corporal y la entonación en el proceso de comunicación. El enfoque de Mehrabian se centró en analizar cómo se transmite la emoción, como el impacto que tendría afirmar palabras con un tono de voz que no coincidiera.

Sus resultados fueron los siguientes: Un 7 % de lo que se transmite en el proceso de la comunicación, consiste en las palabras que decimos, un 38 % viene de la voz, y un 55 % lo transmitimos como comunicación no verbal, es decir, el cuerpo [7].

Michael Kraus, otro psicólogo e investigador, demostró en otro experimento que las señales visuales, es decir, la información no verbal, nos distrae y reduce la capacidad en la detección de emociones. En su estudio, los voluntarios debían identificar las emociones dentro de una conversación con los ojos tapados y sin tapar Cuidando su *voz. (n.d.). NIDCD. Retrieved February 29, 2024, from https://www.nidcd.nih.gov/es/espanol/cuidando-su-voz*

[8]. Con esto, Kraus concluyó que la voz lleva más información emocional y que además es más fácil interpretarla sin señales visuales.

2.3. Recopilación / Referencias de otros estudios

En este apartado se nombran algunos estudios relevantes en el contexto del desarrollo de modelos de detención de emociones en señales de voz. Todos abordan el reconocimiento de emociones en el habla desde diferentes perspectivas y demuestran la viabilidad de utilizar parámetros acústicos para identificar estados emocionales además de explorar el uso de técnicas como Deep learning, análisis de características…

- **Zhang et al. Zhang, B., Essl, G., & Provost, E. M. Emotion from singing and speaking using shared models** [9].

Se trata de uno de los primeros intentos de reconocimiento, emplearon un modelo SVM utilizando ondas de Morlet sobre seis emociones contenidas en la base de datos RAVDESS, logrando una precisión inicial del 42.48%. Este trabajo marcó el punto de partida de futuras investigaciones en este campo

- **Petrushin, V. Emoción en el Habla: Reconocimiento y Aplicación a Call Centers** [10].

Consta de 2 estudios, llevados a cabo por V. Petrushin que trata de enfocarse en el reconocimiento de emociones en el habla con aplicación específica en centros de llamadas. En la primera parte del estudio participan 30 actores no profesionales cuya tarea es interpretar 700 enunciados con cinco emociones distintas: felicidad, ira, tristeza, miedo y estado neutral.

Utiliza el tono de voz, primeros y segundos formantes, energía de la voz y velocidad del habla como características. Se implementó el algoritmo de k-vecinos más cercanos (k-NN) y se logró una precisión de aproximadamente el 70%.

En la segunda fase, 18 actores no profesionales interpretan 56 enunciados con emociones agrupadas en 2 estados: Agitación (incluye ira, felicidad y miedo) y Calma (incluye neutral y tristeza). Con el mismo método de clasificación se alcanzó una precisión del 77%.

- **Nakatsu, R.; Nicholson, J.; Tosa, N. Reconocimiento de emociones y su aplicación a agentes informáticos con capacidades interactivas espontáneas** [11].

Este estudio conducido por R. Nakatsu, J. Nicholson, y N. Tosa, aborda el desafío del reconocimiento de emociones y su aplicación en agentes

informático con habilidades interactivas espontáneas. Se adopta un enfoque basado en redes neuronales:

Se implementa una red neuronal compuesta por 8 subredes, cada una destinada a una emoción concreta. Cada subred consta de 5 capas con 300 nodos de entrada (correspondientes a la dimensión de las características del habla) y 1 nodo de salida.

Como resultado se obtiene una tasa de reconocimiento de aproximadamente 50%.

- **Ashish B. Ingale, D. S. Chaudhari. International Journal of Soft Computing and Engineering** [12]**.**

El estudio también aborda el reconocimiento de emociones en el habla. Diversas características como la energía, el tono, LPCC, MFCC son analizadas para evaluar el rendimiento de diferentes clasificadores. Utilizando GMM (Modelos Gaussianos Mixtos) se obtiene un rendimiento del 75% con un sistema de reconocimiento independiente del hablante y un 89,12% para el reconocimiento dependiente.

Utilizando una red neuronal artificial (RNA) se obtiene un rendimiento del 52,87% para el reconocimiento independiente del hablante.

El uso del Clasificador HMM obtuvo un rendimiento de 76.12% para el reconocimiento del hablante dependiente y un 64,77% para el reconocimiento del hablante independiente.

- **M. Sidorov, S. Ultes, and A. Schmitt, ''Emotions are a personal thing: Towards speaker-adaptive emotion recognition,''** [13]**.**

En mayo de 2014 M. Sidorov, S. Ultes y A. Schmitt. Presentan en la conferencia Internacional de procesamiento de Señales de Audio y Habla un estudio centrado en el reconocimiento de emociones de nuevo con modelos dependientes del hablante. La metodología sigue 2 etapas, donde primero se determina la identidad del hablante para utilizar esta información en el

reconocimiento. La técnica se evaluó en 5 bases de datos en diferentes idiomas y se observó una mejora significativa en la precisión del reconocimiento con un aumento del +10.2% al agregar información específica del hablante.

En resumen, la investigación en reconocimiento de emociones mediante aprendizaje automático ha ido avanzando y explorando una variedad de modelos y técnicas. Aunque los modelos SVM han sido comunes en investigaciones anteriores, el uso de redes neuronales convolucionales CNN y recurrentes RNN continúan ganando cada vez más popularidad.

2.4. Características de las señales de voz

La voz, como sonido que es, se caracteriza por tener una serie de elementos o características esenciales que desempeñan un papel muy importante en el reconocimiento de emociones. Estos componentes del habla son fundamentales, ya que, además de transmitir información explícita, la voz también contiene información acerca de la persona que lo está emitiendo.

Entre las características fundamentales se encuentran el tono, frecuencia fundamental, duración, tono, calidad de voz… todas ellas importantes para entender y expresar emociones. Sin embargo, para implementar un sistema de clasificación, es más útil utilizar características objetivas, es decir, aquellas cuya cuantificación no dependa del criterio del oyente [14].

De este tipo existen dos conjuntos:

- Características prosódicas:

Estas características se centran en cómo decimos las palabras. Incluyen la entonación (cómo sube y baja la voz al hablar), el énfasis en ciertas palabras o las pausas entre frases.

- Características espectrales:

Estas características buscan representar cómo se distribuye la energía en las diferentes frecuencias de la voz. Se centran en cómo suena la voz y es muy útil en por ejemplo la identificación de fonemas, emociones…

Algunos ejemplos son los Coeficientes Cepstrales de frecuencia de Mel, la energía espectral, formantes, entre otros.

La combinación de ambas categorías proporciona una visión más completa en el procesamiento de la señal de voz y ambas son importantes para el reconocimiento de emociones.

Un cambio emocional se ve reflejado en la velocidad del habla, se ve reflejado en el tono o en la forma de onda glótica, energía… Por ejemplo, una persona enojada tiende a hablar con más intensidad y un tono más alto. En un estado de felicidad, también varía el tono y aumenta la energía de la voz.

Por otro lado, la tristeza puede manifestarse con la disminución del tono, la velocidad del habla podría ser más lenta, y el espectro de componentes de alta frecuencia puede disminuir.

Emotions	**Pitch**	**Intensity**	**Speaking rate**	**Voice quality**
Anger	abrupt on stress	much higher	marginally faster	breathy, chest
Disgust	wide, downward inflections	lower	very much faster	grumble chest tone
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflections	higher	faster/slower	breathy, blaring tone
Joy	high mean, wide range	higher	faster	breathy; blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

Tabla 2 Relación emoción – característica [15].

Es importante destacar que la relación entre las características de la voz y la propia emoción (Tabla 2) puede variar entre hablantes, estilos de expresión y culturas, como se viene diciendo en apartados anteriores. Es por esto por lo que,

no existe una fórmula exacta para distinguir todas las emociones igual tampoco existe un conjunto definitivo de características acústicas que permita distinguir todas las emociones.

3. ESTUDIO TEÓRICO

3.1. Inteligencia artificial

Remontémonos al siglo XIX para introducir el concepto de inteligencia artificial. En 1854, el matemático George Boole presentó en su obra '*An Investigation of the Laws of Thought*' una nueva lógica que podría ser sistematizada utilizando un sistema algebraico, este trabajo sentó las bases para entender la lógica de manera más matemática [16].

No fue hasta 1936 cuando el visionario Alan Turing [17], mientras ya se teorizada sobre 'máquinas' capaces de funcionar de manera autónoma, introdujo el concepto de algoritmo. Esta contribución allanó el camino para el surgimiento de la informática moderna, construyendo la base sobre la cual se apoyarían las futuras investigaciones de inteligencia artificial.

McCarthy, considerado 'el padre de la inteligencia artificial', fue quien organizó la famosa conferencia de Dartmouth en 1956 y junto a Marvin Minsky y Claude Shannon comienzan a desarrollar las primeras teorías y modelos en IA. Planteaban una sociedad repleta de máquinas en menos de 10 años, sin embargo, la falta de recursos y avances prácticos no dejaron avanzar la IA hasta la década de los 90.

El año 1997 se considera un punto de inflexión en la historia de la inteligencia artificial. Una reconocida empresa estadounidense llamada IBM lanza Deep Blue, una supercomputadora diseñada para jugar al ajedrez y competir contra el campeón mundial de ajedrez, Garry Kaspárov [18]. La victoria de *Deep Blue* demostró la capacidad de las máquinas para participar en tareas humanas.

Este hito sirvió como resurgimiento en la investigación de la inteligencia artificial, y figuras como Geoffrey Hinton, Yann LeCun o Yodhua Bengio

comenzaron a desarrollar algoritmos y arquitecturas de redes neuronales, dotando a las máquinas de capacidad de comprensión y respuesta.

Desde entonces, los sistemas van adquiriendo nuevas habilidades, como la capacidad de cálculo, de almacenar información… sin embargo, existen tareas que, hasta el momento se consideran 'demasiado humanas' para una máquina. Capacidades como la creatividad, innovación, autoconciencia o transmisión de emociones aún quedan fuera de su alcance.

Actualmente, la IA se define como una disciplina que combina algoritmos con el objetivo de desarrollar máquinas que imiten la habilidad cognitiva de los seres humanos, proporcionando soluciones a problemas complejos. Esto implica realizar un esfuerzo por construir sistemas capaces de aprender de su entorno, de los errores y de las personas.

3.1.1. Aprendizaje automático o Machine Learning

Dentro del amplio espectro de la inteligencia artificial, se encuentra un subconjunto conocido como aprendizaje automático (ML por sus siglas en inglés), que comprende las técnicas y algoritmos que permiten que un sistema adquiera conocimiento por sí mismo. En este proceso de aprendizaje, el sistema utiliza un conjunto de datos de entrada y a partir de estos, busca patrones que le capaciten para tomar decisiones, aprender de errores pasados y realizar predicciones sobre futuros escenarios.

El aprendizaje automático se asemeja al proceso de aprendizaje humano, dividiéndose en dos fases fundamentales: la fase de entrenamiento y la fase de prueba. Durante la primera, el sistema asimila la información extraída del conjunto de datos y desarrolla una comprensión de los patrones presentes en dichos datos. En la segunda, se evalúa la capacidad del modelo para utilizar ese conocimiento en situaciones nuevas y hacer predicciones precisas.

Se distinguen tres tipos de aprendizaje según la naturaleza de los datos y la forma en la que las máquinas adquieren la información [19]:

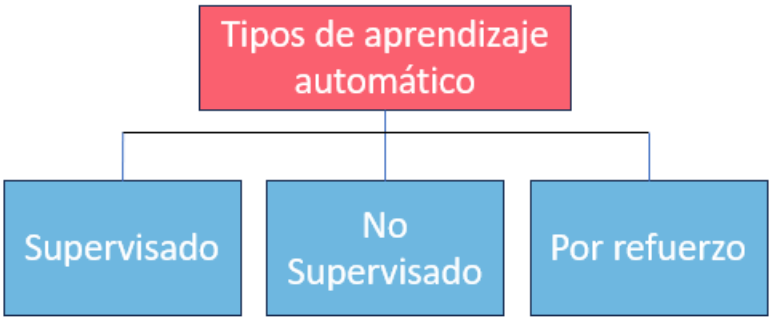


Figura 4. Tipos de aprendizaje de máquina

- **Aprendizaje supervisado**

Este algoritmo se entrena a partir de datos debidamente etiquetados, se establece una correspondencia entre los datos de entrada (x, *inputs*) y los datos de salida (y, *outputs*), permitiendo al modelo aprender a predecir resultados. Se requieren estos conjuntos de datos de entrenamiento y de prueba, los primeros para el entrenamiento y los segundos para determinar cuánto de eficaz es el modelo [20].

Se denomina 'supervisado' porque ya se conoce la respuesta deseada. En este proyecto se utilizan datos extraídos de una base de datos previamente etiquetados y categorizados en diferentes emociones, con ayuda de intervención humana.

Dentro de este aprendizaje se distinguen las técnicas de clasificación y regresión. El algoritmo de clasificación etiqueta cada muestra eligiendo entre diferentes clases dependiendo de si se trata de una clasificación binaria o multiclase. La regresión, predice valores continuos basados en las entradas anteriores.

Se detalla a continuación el proceso que sigue el aprendizaje supervisado:

- Fase de entrenamiento

En una primera etapa, el sistema adquiere las muestras extraídas de una o varias bases de datos existentes y sin procesar. El siguiente paso es etiquetar correctamente estos datos de entrenamiento, creando un vector m x 1 para asociar las entradas con las salidas deseadas, siendo m el número de ejemplos

que se utilizan. A continuación, se extraen características de las instancias etiquetadas, formando una matriz m x n que mapea cada instancia en un espacio n-dimensional, siendo n el número de características extraídas. Para esta tarea se utilizan clasificadores que generan un modelo que discrimine de manera eficiente las instancias según sus etiquetas.

- Fase de evaluación o prueba

Se utilizan instancias no procesadas para evaluar el modelo en condiciones no vistas previamente, es muy importante que las muestras no hayan sido utilizadas durante el entrenamiento. Estas nuevas muestras se etiquetan generando un vector mp x 1 con los resultados de la categoría asignada.

Cabe destacar que, durante ambas fases, los clasificadores que el sistema utiliza para aprender aplican funciones que permiten diferenciar y clasificar eficientemente el modelo. Son una herramienta clave a la que se dedicará un apartado en esta memoria.

- **Aprendizaje no supervisado**

Esta modalidad de aprendizaje se destaca por su capacidad para no depender de datos etiquetados durante el proceso de entrenamiento. El sistema, ajeno a la intervención humana, aprende de datos sin etiquetar y toma sus propias decisiones descubriendo nuevos patrones. Dado que no hay etiquetas en el proceso, la evaluación de resultados se vuelve un reto, ya que no existe una medida directa de precisión.

El aprendizaje no supervisado se utiliza, por ejemplo, para automatizar recomendaciones de música. Al carecer de etiquetas, es especialmente útil para encontrar similitudes en las preferencias musicales de los usuarios. Agrupa de forma automática a usuarios con gustos similares y genera recomendaciones personalizadas basadas en las elecciones de grupos afines.

- **Aprendizaje por refuerzo**

En este paradigma, el sistema, también llamado agente, aprende en un entorno interactivo mediante prueba y error. En este proceso, el agente no recibe instrucciones, sino que debe descubrir por sí mismo qué acciones conducen a la máxima recompensa. Se recibirá una respuesta positiva o negativa en función del comportamiento del sistema.

Un ejemplo ilustrativo de este tipo de aprendizaje es una prótesis inteligente de una extremidad. El modelo realiza intentos de movimiento sin orientación previa y recibe recompensa cuando esos movimientos 'a ciegas' son efectivos. Es un aprendizaje semi – supervisado porque fusiona los dos tipos de aprendizaje estudiados anteriormente.

3.1.2. Aprendizaje profundo o *Deep Learning*

Se trata de un subconjunto del *Machine Learning* que se centra en el uso de redes neuronales profundas, a través de estas, el *Deep Learning* imita el procesamiento cerebral utilizando algoritmos que conforman redes neuronales con múltiples capas interconectadas.

Igualmente, el aprendizaje profundo destaca por su capacidad para manejar grandes volúmenes de datos durante el entrenamiento, por lo que su capacidad de predicción es más precisa.

El Deep Learning se utiliza en la identificación de objetos, en el reconocimiento en señales de audio, traducción de idiomas… al estar estructurado de manera similar al cerebro humano, permite que la IA aprenda de manera continua y mejore a medida que se le proporciona más información.

Mientras que el *Machine Learning* depende en gran medida de la selección manual de características, el Deep Learning utiliza redes neuronales profundas para aprender automáticamente representaciones más complejas y no requiere trabajar con datos ordenados, etiquetados y estructurados, es decir, logra una mayor autonomía [21].

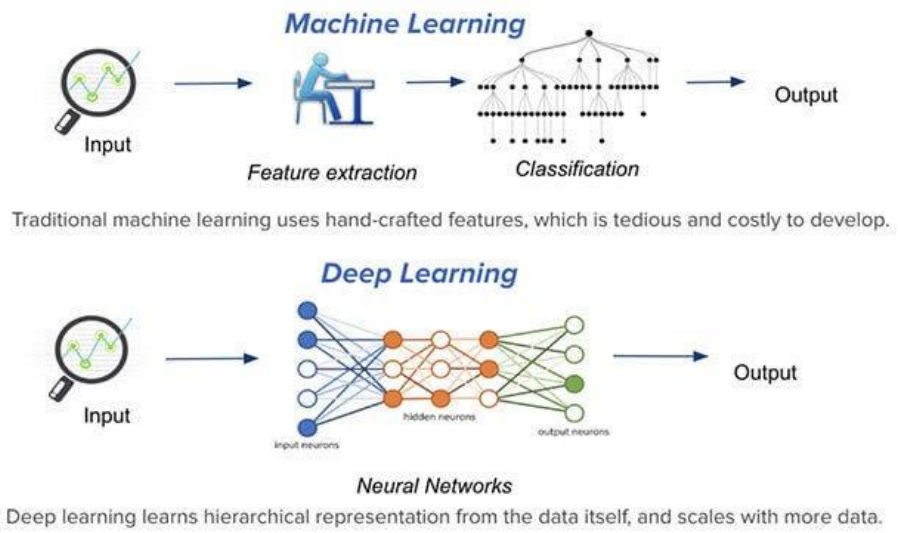


Figura 5. Comparativa Machine Learning vs Deep Learning [22].

3.2. Redes neuronales (NN) artificiales (ANN)

Las redes neuronales forman parte del aprendizaje automático y son fundamentales para el aprendizaje profundo. Se trata de un modelo computacional inspirado en el cerebro humano que trata de imitar la manera en la que nuestras neuronas se relacionan entre sí.

Consiste en un conjunto de nodos, denominados neuronas artificiales, organizados en capas. Estas capas incluyen la capa de entrada, donde la red recibe información; una o más capas ocultas; y la capa de salida, donde la red proporciona la respuesta final [23].

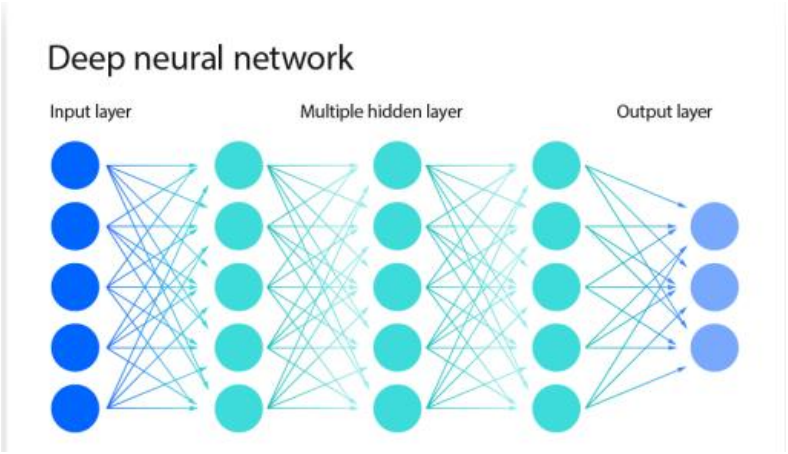


Figura 6. Ejemplo de una red neuronal con 3 capas ocultas [23]

A continuación, se ve más detalladamente el funcionamiento de una red neuronal:

Una vez que se establece una capa de entrada, se asignan ponderaciones que indican la importancia relativa de cada variable de entrada, los valores más grandes significan una contribución más significativa a la salida.

En este ejemplo (Figura 7) la neurona tiene 5 entradas identificadas como x1, x2, x3, x4, x5. Cada una de estas entradas se conecta por medio de cinco enlaces con un peso asociado denominado w1, w2, w3, w4, w5 [23].

El proceso consiste en multiplicar cada valor de entrada por su respectivo peso, la suma de estos productos da como resultado la entrada total de la neurona.

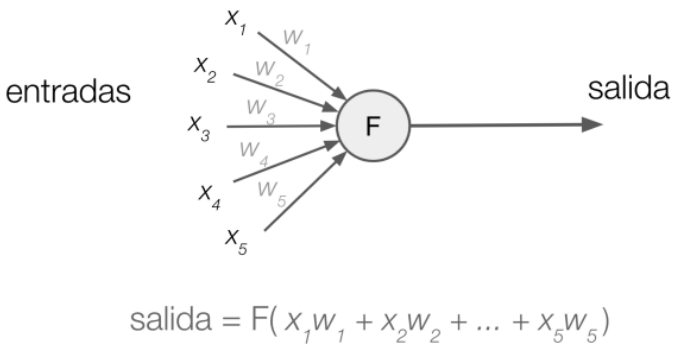


Figura 7. Descripción neurona con 5 entradas [24].

El conjunto de sumas de todas las entradas se pasa a través de una función de activación, que determina la salida final.

Si esta salida super cierto umbral, el nodo se activa y pasa los datos a la siguiente capa de la red neuronal convirtiéndose en la entrada del siguiente nodo.

Antes de profundizar en la clasificación de redes neuronales, vamos a definir la nombrada 'función de activación'.

Se trata de una especie de filtro que procesa la combinación de pesos y entradas que se han mencionado. Su función es transmitir la información a través de las capas de la red. Existen varias, aquí se muestras las más comunes:

- **Función ReLU**: Transforma los valores de entrada anulando los negativos y dejando intactos los positivos.
- **Función Sigmoide**: Transforma los valores de entrada a una escala entre 0 y 1. Los valores altos tienden a 1 mientras que los valores bajos tienden a 0.
- **Función Tangente Hiperbólica**: Transforma los valores de entrada a una escala entre -1 y 1. Al igual que la función sigmoide, los valores altos tienden a 1 mientras que los valores bajos tienden hacia -1.

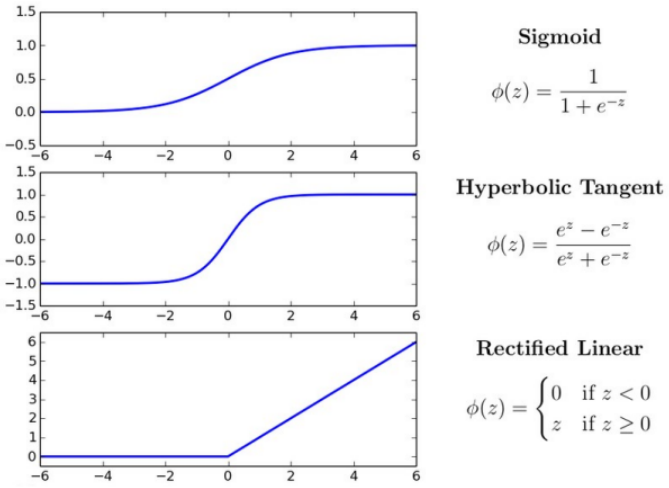


Figura 8. Representación y fórmula de las funciones de activación más comunes [24].

3.2.1. Redes Neuronales de alimentación hacia adelante *(Feedforward)* o perceptrones multicapa MPL

Estas redes se componen de una capa de entrada, una o varias capas intermedias y una capa de salida. Cada nodo en una capa se conecta a los nodos de la siguiente mediante conexiones ponderadas con pesos. La información se mueve en una sola dirección, desde la capa de entrada hacia

la de salida, sin bucles ni retroalimentación directa. Este diseño facilita el entrenamiento y la interpretación del modelo.

Las MPLs son eficaces para problemas de clasificación y regresión en los que la relación entre las características de entrada y las salidas es relativamente simple y no depende del orden temporal o de secuencia.

Se ha visto que la predicción para un proceso de regresión lineal es el siguiente:

$$\hat{Y} = w[0] * x[0] + w[1] * x[1] + \cdots + w[n] * x[n] + b, \quad (1)$$

siendo b el sesgo que permite que la línea de regresión tenga una inclinación y una posición vertical adecuadas para ajustarse mejor a los datos. Representa el valor de y cuando todas las características de entrada son iguales a 0 [26].

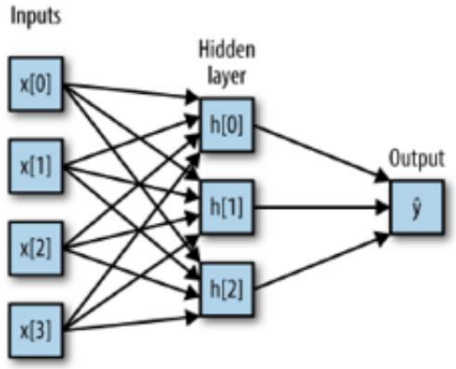


Figura 9. Arquitectura básica red neuronal [27].

Cada nodo de la izquierda x(n) representa una característica, estas se conectan a través de líneas que representan los pesos asociados a las características, con la salida representada al lado derecho y es resultado del sumatorio de pesos y entradas.

En una red neuronal de tipo MPL, este proceso se repite en múltiples capas, convirtiéndose cada salida en la entrada de otra capa (oculta). Finalmente, después de varias capas ocultas, se obtiene un sumatorio al que como se ha explicado antes, se le aplica una función no lineal (ReLU, tangente hiperbólica, etc.)

3.2.2. Redes Convolucionales (CNN)

Este modelo de red neuronal es común por su gran rendimiento en el procesado de imágenes y vídeo ya que son redes que están diseñadas para trabajar con datos de entrada dispuestos en una especie de cuadrícula, como las matrices. Estas redes procesan la entrada en varias etapas, culminando en una ANN que actúa como un clasificador.

Se componen de 3 tipos de capas [28]:

- Capa convolucional *(convolution layer)*

Es la primera capa donde comienza a crear la red y donde se realizan la mayoría de los cálculos utilizando un filtro conocido como Kernel para detectar características específicas en la imagen de entrada, como bordes, colores o texturas. Este proceso se denomina convolución y tiene como resultado un nuevo conjunto de datos llamado 'mapa de características' o 'mapa de activación' donde se muestran las características importantes de la imagen.

Sobre esta salida se aplica una función ReLU para introducir la no linealidad al modelo, para así aumentar la probabilidad de que la red aprenda patrones más complejos durante el entrenamiento.

- Capa de agrupación *(polling layer)*

Conocida como submuestreo, es una operación cuya función principal es reducir la cantidad de información en los mapas de características obtenidos en las capas anteriores.

Esta reducción consiste en agrupar la información utilizando un filtro, en este caso que no utiliza pesos, que se mueve sobre la entrada sin modificar sus valores.

Disminuyendo la cantidad de información, se simplifica el procesamiento en las capas posteriores, lo que implicará menos cálculos y, por lo tanto, un modelo

más eficiente. Además, al conservar menos detalles, se limita el riesgo de que la red memorice el conjunto de entrenamiento y prevé el sobreajuste.

- Capa totalmente conectada *(fully connected layer)*

Es la parte de la red que toma todas las características aprendidas hasta el momento y las utiliza para la tarea final de clasificación, asignando probabilidades a las posibles categorías y seleccionando la más probable como la predicción final.

En este caso, cada nodo de la capa de salida está vinculado directamente a cada nodo de la capa anterior, basándose en las características que la red ha aprendido en capas anteriores, decide a qué categoría pertenece la imagen. En nuestro caso, esta capa decidiría si la imagen representa la emoción triste o enfado.

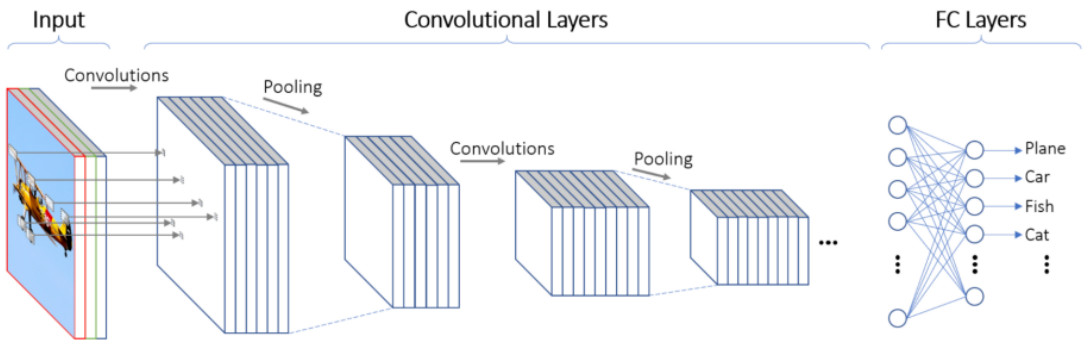


Figura 10. Esquema CNN [29].

3.2.3. Redes Recurrentes (RNN)

Son muy buenas para trabajar con datos que tienen un orden específico, como palabras en una oración, pasos en una receta… muy útiles por lo tanto en el procesamiento de audio, en la tarea de agregar subtítulos a imágenes y en la traducción de idiomas, entre otros.

Como los modelos anteriores, a partir de muchos ejemplos de datos secuenciales, la RNN aprende a reconocer patrones y hacer predicciones.

Se caracteriza por su memoria, este tipo de red puede recordar cosas del pasado. A medida que la red avanza, obtiene información de las anteriores neuronas y de ella misma en el paso previo. Además, a diferencia de otras redes neuronales, no asume que cada pieza de información es independiente si no que su salida depende de lo que ha sucedido antes en la secuencia [30].

Otra característica distintiva de este modelo es que comparten el mismo conjunto de pesos en cada capa de red, aunque los pesos sean comunes en cada capa, igualmente, estos se ajustan durante los procesos de retropropagación y pendiente de gradiente para mejorar el aprendizaje.

Utilizan el algoritmo de retropropagación a través del tiempo (BPTT) para determinar los gradientes. Aunque sigue los mismos principios que la retropropagación tradicional, la BPTT tiene un paso adicional en el que suma los errores en cada paso temporal para poder compartir los pesos en cada capa.

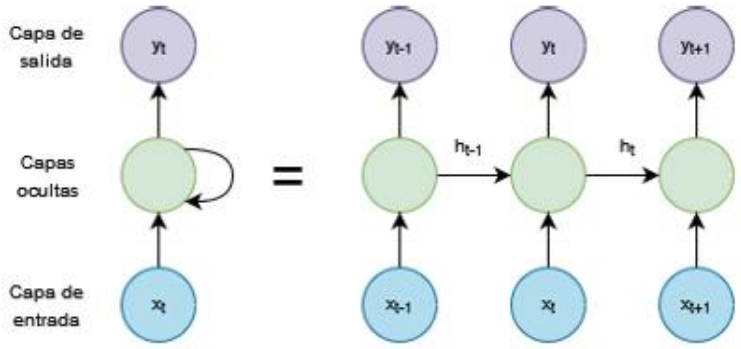


Figura 11. Representación de una red neuronal recurrente [30].

Dentro de este tipo se pueden presentar dos problemas relacionados con los gradientes, que son los indicadores de qué tan rápido está aprendiendo el modelo:

- *Gradients Vanishing*: Cuando los gradientes son demasiado pequeños, los parámetros de peso tienden a 0 y se detiene el aprendizaje
- *Gradients Exploding*: Cuando los gradientes son demasiado grandes, se crea un modelo inestable con el crecimiento exponencial de los pesos.

La solución a este problema, que, aunque menos comúnmente, también aparece en redes MLP, se presenta durante mi entrenamiento, se presentará en la última parte de la memoria junto al resto de análisis de resultados.

Por último, y al igual que en otros tipos de redes, se utilizan funciones de activación para introducir no linealidades en el modelo y permitir que la red aprenda patrones más complejos.

3.2.4. Redes Neuronales de Memoria a Corto y Largo Plazo (LSTM)

Se trata de una variante de red Neuronal recurrente tiene la capacidad de decidir qué información retener y qué olvidar, está diseñada para recordar la información importante durante mucho tiempo y olvidar de manera más efectiva, lo que la hace útil en tareas que requieren compresión a largo plazo, como el procesamiento de lenguaje natural. Fue creada para superar las limitaciones con los gradientes de las RNN.

Las LSTMs tienen tres puertas principales que controlan el flujo de información dentro de la celda de memoria: La puerta de olvido decide qué información almacenada en la celda debe olvidarse; la puerta de entrada indica qué información debe agregarse a la celda y la puerta de salida decide qué información de la celda de memoria se utilizará como salida [31]

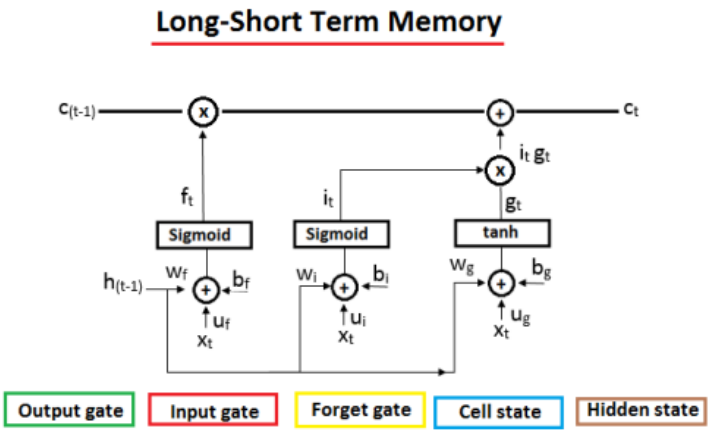


Figura 12. Ejemplo de una red LSTM [32].

3.3. Computación afectiva

La computación afectiva es una rama dentro de la IA desde 1995, también conocida como Inteligencia Artificial Emocional. Esta 'reciente' disciplina estudia cómo diseñar máquinas que puedan reconocer, interpretar y dar una respuesta apropiada a las emociones humanas adaptando su comportamiento en base a ese conocimiento.

Esta disciplina utiliza el aprendizaje automático para extraer las características más importantes de los datos (imágenes, voces, etc.) y posteriormente procesarlas. Con esto, la computación afectiva busca emular la interacción entre personas, combinando el lenguaje no verbal (postura, expresividad) con el verbal.

Los ordenadores no pueden sentir, sin embargo, a través de técnicas como Deep learning, se han logrado importantes avances en el reconocimiento de emociones, lo que ha permitido su implementación en diversos campos, siempre con el fin de responder correctamente ante nuestras peticiones [33][34].

3.3.1. Aplicaciones de la computación afectiva

Los sistemas de identificación de emociones tienen una amplia e interesante gama de aplicaciones en la actualidad y que está en continuo desarrollo para mejorar la interacción entre humanos y máquinas y enriquecer nuestras experiencias cotidianas [34]

Aplicaciones que contribuyen al bienestar y comodidad de las personas:

- En los '*Call Centers*" y servicios al cliente: Se utilizan sistemas automatizados para detectar emociones y adaptar las respuestas a la persona que llama, o incluso transferir la llamada a un agente humano [34]

- Asistentes de voz: En dispositivos con inteligencia artificial, el reconocimiento de emociones ayuda a comprender mejor las necesidades y mejora la interacción y las respuestas de los asistentes virtuales.

- Industria automovilística: se pueden detectar señales de fatiga y distracción en el conductor para mejorar la seguridad vial

- Salud mental y bienestar: En el ámbito de la salud, es beneficioso controlar los estados emocionales de las personas con trastornos mentales y proporcionar apoyo emocional y atención personalizada.

- *Marketing*: El reconocimiento de emociones en las respuestas de los clientes es una herramienta valiosa para personalizar estrategias publicitarias y medir la efectividad de campañas emocionales.

- Industria del entretenimiento: se mejora la experiencia del usuario adaptando los juegos y aplicaciones en función de sus emociones.

Aplicaciones y tecnologías existentes [35]

- Asistentes virtuales como *Siri, Alexa* y *Google Assistant* emplean el reconocimiento de emociones para proporcionar respuestas más humanas y personalizadas.

- *Moodnotes*: se trata de una aplicación de bienestar mental que permite al usuario realizar un seguimiento de sus emociones. Utiliza el reconocimiento para ayudar a los usuarios a identificar patrones emocionales y mejorar su inteligencia emocional.

- *Replika:* La aplicación se centra en crear una experiencia de conversación más humana y emocional mediante un *chatbot.*

- *Affectiva:* Esta empresa ofrece una API de reconocimiento de emociones que puede integrarse en, por ejemplo, la medición de emociones en entornos de conducción [34].

3.4. Bases de datos

Una base de datos es una colección de datos estructurados que abarca formatos como imágenes, vídeos, voz… Para este proyecto se busca un

conjunto de audios cuyo contenido se corresponde con expresiones de emociones.

En el ámbito de la Inteligencia Artificial, los datos son esenciales, se puede decir que son el combustible que alimenta el desarrollo y el entrenamiento de modelos. La IA ya recopila, agrega, procesa y gestiona volúmenes muy grandes de datos.

Existen muchas bases de datos de emociones, cada una con sus propias limitaciones. Factores como la calidad de las grabaciones, el número y diversidad de emociones, la cantidad de participantes o el idioma, entre otros, condicionan la tasa de reconocimiento.

La elección del conjunto de datos para alimentar al sistema plantea la posibilidad de incorporar emociones auténticas o emociones actuadas. El modelo será más robusto con el uso de una base de datos recopilada en situaciones de la vida real, ya que el reconocimiento de patrones en emociones reales se alinea mejor con el entorno para el cual se está desarrollando un modelo.

Los conjuntos de datos, específicamente en el ámbito de reconocimiento de voz, se pueden distinguir 3 categorías:

- Colección de datos de habla naturales: provienen de fuentes como la televisión o internet. Detectar y modelar estos conjuntos es complicado debido a la continua variación dinámica de las emociones durante la emisión del sonido. Además, es difícil recoger este tipo de datos por los posibles problemas relacionados con derechos de autor y privacidad.
- Colección de datos seminaturales: estos conjuntos se generan a través de la representación de actores de voz que simulan escenarios emocionales. Aun que comparten similitudes con expresiones naturales, no dejan de estar creadas artificialmente.
- Colección de datos simulados: Similar al conjunto anterior, pero en este caso, los actores dictan las mismas sentencias con emociones específicas. En lugar de representar escenarios, se expresa una gama predefinida de emociones al decir las mismas frases. Estas colecciones permiten una comparación directa de los resultados entre diferentes

modelos y enfoques, sin embargo, presentan la desventaja de inclinarse a modelos sobre ajustados.

Corpus	Access	Language	Size	Source	Emotions
LDC Emotional Prosody Speech and Transcripts	Commercially available[a]	English	7 actors × 15 emotions × 10 utterances	Professional actors	Neutral, panic, anxiety, hot anger, cold anger, despair, sadness, elation, joy, interest, boredom, shame, pride, contempt
Berlin emotional database	Public and free[b]	German	800 utterances (10 actors × 7 emotions × 10 utterances + some second version) = 800 utterances	Professional actors	Anger, joy, sadness, fear, disgust, boredom, neutral
Danish emotional database	Public with license fee[c]	Danish	4 actors × 5 emotions (2 words + 9 sentences + 2 passages)	Nonprofessional actors	Anger, joy, sadness, surprise, neutral
Natural	Private	Mandarin	388 utterances, 11 speakers, 2 emotions	Call centers	Anger, neutral
ESMBS	Private	Mandarin	720 utterances, 12 speakers, 6 emotions	Nonprofessional actors	Anger, joy, sadness, disgust, fear, surprise
INTERFACE	Commercially available[d]	English, Slovenian, Spanish, French	English (186 utterances), Slovenian (190 utterances), Spanish (184 utterances), French (175 utterances)	Actors	Anger, disgust, fear, joy, surprise, sadness, slow neutral, fast neutral
KISMET	Private	American English	1002 utterances, 3 female speakers, 5 emotions	Nonprofessional actors	Approval, attention, prohibition, soothing, neutral
BabyEars	Private	English	509 utterances, 12 actors (6 males + 6 females), 3 emotions	Mothers and fathers	Approval, attention, prohibition
SUSAS	Public with license fee[e]	English	16,000 utterances, 32 actors (13 females + 19 males)	Speech under simulated and actual stress	Four stress styles: Simulated Stress, Calibrated Workload Tracking Task, Acquisition and Compensatory Tracking Task, Amusement Park Roller-Coaster, Helicopter Cockpit Recordings
MPEG-4	Private	English	2440 utterances, 35 speakers	U.S. American movies	Joy, anger, disgust, fear, sadness, surprise, neutral
Beihang University	Private	Mandarin	7 actors × 5 emotions × 20 utterances	Nonprofessional actors	Anger, joy, sadness, disgust, surprise
FERMUS III	Public with license fee[f]	German, English	2829 utterances, 7 emotions, 13 actors	Automotive environment	Anger, disgust, joy, neutral, sadness, surprise
KES	Private	Korean	5400 utterances, 10 actors	Nonprofessional actors	Neutral, joy, sadness, anger
CLDC	Private	Chinese	1200 utterances, 4 actors	Nonprofessional actors	Joy, anger, surprise, fear, neutral, sadness
Hao Hu et al.	Private	Chinese	8 actors × 5 emotions × 40 utterances	Nonprofessional actors	Anger, fear, joy, sadness, neutral
Amir et al.	Private	Hebrew	60 Hebrew and 1 Russian actors	Nonprofessional actors	Anger, disgust, fear, joy, neutral, sadness
Pereira	Private	English	2 actors × 5 emotions × 8 utterances	Nonprofessional actors	Hot anger, cold anger, joy, neutral, sadness

Tabla 3. Comparativa bases de datos conocidas.

3.5. Métricas de evaluación

Esta sección tiene un papel muy importante en la evaluación del rendimiento de la red neuronal que se desarrolla en el proyecto. La eficacia y fiabilidad de cualquier sistema de reconocimiento de emociones se mide mediante métricas que cuantifican su capacidad de clasificar con precisión los diferentes estados emocionales [38][39].

1. Matriz de confusión:

La matriz de confusión proporciona una visión detallada del rendimiento del modelo, pues permite visualizar el desempeño de un modelo de clasificación al mostrar cuándo una clase es confundida con otra desglosando las predicciones en [38]:

- Verdaderos positivos (TP): instancias positivas correctamente identificadas como positivas. (Persona feliz, el modelo lo clasifica como feliz).
- Falsos positivos (FP): instancias negativas incorrectamente identificadas como positivas. (Persona no feliz, el modelo lo clasifica como feliz).
- Verdaderos negativos (TN): instancias negativas correctamente identificadas como negativas. (Persona no feliz, el modelo lo clasifica como no feliz).
- Falsos negativos (FN): instancias positivas incorrectamente identificadas como negativas. (Persona feliz, el modelo lo clasifica como no feliz).

Una matriz de confusión tiene la siguiente forma:

TP FP

FN TN

Esta métrica sirve para posteriormente evaluar la precisión, sensibilidad y otras métricas que se mencionan a continuación.

2. Exactitud (*Accuracy*) [39]:

Cantidad total de instancias correctamente identificadas.

$$Accuracy = \frac{\text{Verdaderos Positivos} + \text{Verdaderos Negativos}}{\text{Total de instancias}} \quad (2)$$

3. Precisión (*Precision*) [39]:

Indica la proporción de instancias positivas que son correctamente identificadas como positivas.

En la práctica se refiere al porcentaje de casos positivos detectados.

$$Precision = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Positivos}} \quad (3)$$

4. Sensibilidad (*Recall*) [39]:

Se refiere a la proporción de instancias positivas correctamente clasificadas por el modelo. Es decir, mide la capacidad del modelo para distinguir correctamente las instancias positivas y usa la siguiente fórmula.

$$Recall = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}} \quad (4)$$

5. Especificidad (*Especificity*) [39]:

Se refiere a la proporción de instancias negativas correctamente clasificadas por el modelo. Es decir, mide la capacidad del modelo para distinguir correctamente las instancias negativas con la siguiente fórmula.

$$Especificity = \frac{\text{Verdaderos Negativos}}{\text{Verdaderos Negativos} + \text{Falsos Positivos}} \quad (5)$$

6. *F1-Score*:

Proporciona un equilibrio combinando la precisión y el Recall en una métrica, es muy útil para detectar un desequilibrio entre clases (hay más muestras con la emoción triste que con la emoción neutral).

$$F1\ Score = \frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{ Recall}} \quad (6)$$

7. Curva ROC y Área bajo la Curva ROC (AUC-ROC) [39]:

Esta herramienta representa gráficamente la relación entre la sensibilidad (verdaderos positivos) y la especificidad (falsos positivos), la métrica asociada a la Curva Roc es el Área bajo la curva, esta medida muestra un valor numérico para evaluar el rendimiento. Un AUC-ROC cercano a 1 indicaría un buen rendimiento mientras que un valor cercano a 0.5 sugiere un rendimiento similar al azar (en este caso, en problemas de clasificación binaria).

La curva ROC es una representación visual como menciono en el párrafo anterior, sin embargo, el AUC-ROC se calcula como la integral bajo la curva ROC[40].

$$AUC - ROC = \int Curva\ ROC\ d(Tasa\ de\ Falsos\ Positivos) \tag{7}$$

8. Validación Cruzada (*Cross-Validation*):

Es otra técnica de evaluación del rendimiento de un modelo de aprendizaje, divide el conjunto de datos en múltiples subconjuntos de entrenamiento y prueba varias veces de manera que el modelo se evalúa en diferentes particiones de datos. La validación cruzada proporciona robustez, especialmente cuando el tamaño del conjunto de muestras es limitado.

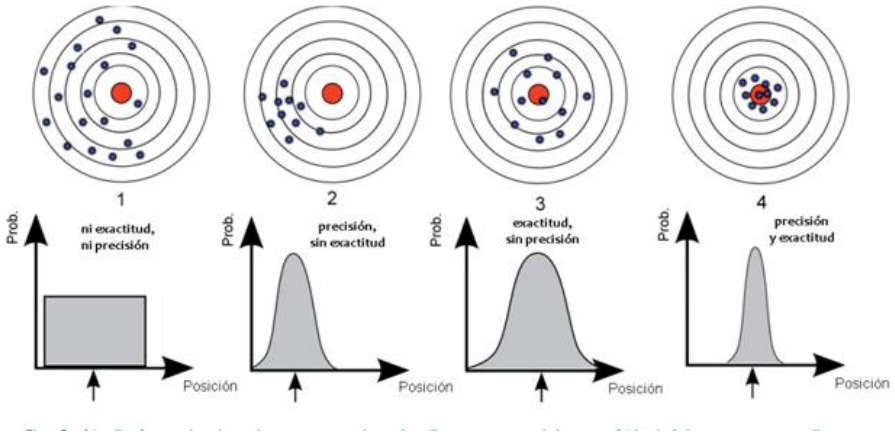


Figura 13. Comparación precisión y exactitud [41].

4. DESARROLLO DEL PROYECTO

4.1. Base de datos utilizada

RAVDESS

La base de datos RAVDESS [1](*Ryerson Audiovisual Database of Emotional Speech and Song*) es una colección multimodal de datos libre interpretada por actores profesionales, en total, contiene 7356 muestras que incluyen grabaciones de habla, canciones y contenido audiovisual [42].

En el contexto de mi investigación en el procesamiento de voz y emociones, he seleccionado los 1440 archivos que corresponden a los audios. Estos archivos están etiquetados con detalles sobre el actor, genero, edad, emoción y tipo de grabación (voz o canción). Las grabaciones realizadas en Toronto, Canadá, fueron interpretadas por 24 actores de ambos géneros (12 hombre y 12 mujeres), quienes expresan emociones en dos niveles de intensidad emocional (normal y fuerte), junto con una expresión neutral. La base de datos incluye las siguientes emociones: neutral, calma, felicidad, tristeza, enojo, miedo, disgusto y sorpresa. Cada grabación tiene una duración de entre 3 y 5 segundos, con un breve silencio inicial de aproximadamente 0.5 segundos y una frecuencia de muestreo de 48kHz.

Cada archivo en la base de datos RAVDESS posee un nombre único que refleja un identificador numérico de 7 partes, definiendo características como modalidad, canal vocal, emoción, intensidad emocional, enunciado, repetición y actor. Por ejemplo, el nombre de un archivo podría ser 03-01-06-01-02-01-12.mp4.

Teniendo en cuenta la nomenclatura que siguen:

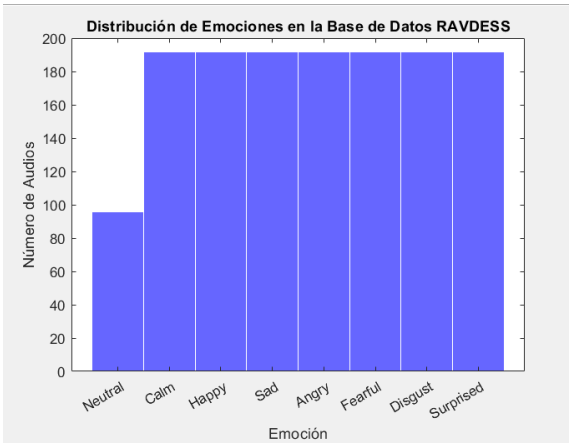
- Modalidad (01: audio-vídeo, 02: sólo vídeo, 03: sólo audio). Canal vocal (01: discurso, 02: canción).
- Emoción (01: neutro, 02: calmado, 03: alegre, 04: triste, 05: enfadado, 06: temeroso, 07: disgustado, 08: sorprendido).

[1] https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

- Intensidad emocional (01: normal, 02: fuerte).
- Enunciado (0“: "Kids are talking by the d”or", 0“: "Dogs are sitting by the d”or").
- Repetición (01: 1º repetición, 02: 2º repetición).
- Actor (01 a 24. Los actores impares son hombres y los pares son mujeres).

Indicaría lo siguiente:

- Archivo de solo audio (03)
- Discurso (01)
- Expresión temerosa (06)
- Intensidad emocional normal (01)
- Enuncia“o "*Dogs are sitting by the d”or*" (02)
- 1ª repetición (01)
- Interpretado por el actor número 12 (12)
- Mujer (ya que el número de identificación del actor es par)




```
Número de muestras por emoción:
Neutral: 96
Calm: 192
Happy: 192
Sad: 192
Angry: 192
Fearful: 192
Disgust: 192
Surprised: 192
```

Figura 14. Visualización distribución de Emociones en la base de datos RAVDESS. (a) Gráfica de la distribución. (b) Número de muestras por emoción.

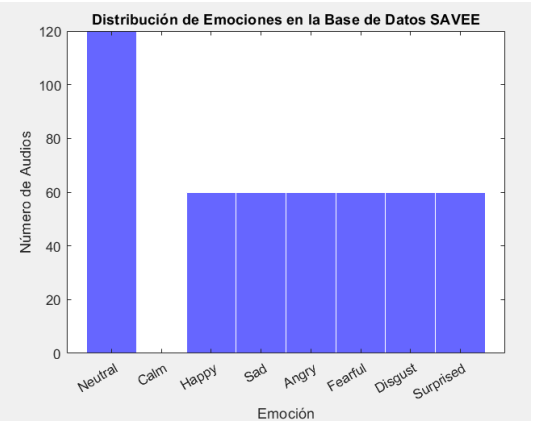
SAVEE

La base de datos *Surrey Audiovisual Expressed Emotion* (SAVEE)[2] se ha registrado a partir de la contribución de cuatro hablantes nativos de inglés, todos ellos de género masculino, identificados como DC, JE, JK y KL. Estos hablantes, estudiantes de posgrado e investigadores de la Universidad de Surrey, tienen entre 27 y 31 años. SAVEE aborda seis distintas emociones: ira, disgusto, miedo, alegría, tristeza y sorpresa. Además, se ha agregado una séptima emoción categoría neutral.

Por cada emoción se identifican 15 oraciones: 3 comunes, 2 específicas para cada emoción y 10 genéricas, las cuales son distintas para cada emoción y cuidadosamente equilibradas fonéticamente. Por otra parte, se registran 30 oraciones neutrales que resultan en 120 expresiones [43] .

Cada archivo de SAVEE consta de un identificador único de 3 partes (por ejemplo, DC_a11.wav).

1. ID del orador (DC, JE, JK y KL).
2. Emoción (a: enfado, d: disgusto, f: miedo, h: alegría, n: neutro, sa: tristeza, su: sorpresa).
3. Oración (01-30).




```
Número de muestras por emoción en SAVEE:
Neutral: 120
Calm: 0
Happy: 60
Sad: 60
Angry: 60
Fearful: 60
Disgust: 60
Surprised: 60
```

Figura 15. Visualización distribución de Emociones en la base de datos SAVEE. (a) Gráfica de la distribución. (b) Número de muestras por emoción.

[2] https://www.kaggle.com/datasets/barelydedicated/savee-database

CREMA-D

Crowd-sourced Emotional Mutimodal Actors Dataset (Crema-D) [3]. Se trata de una rica colección de datos que engloba 7.442 archivos originales, interpretados por 91 actores con edades que varían entre los 20 y 74 años, y provienen de diversas razas y etnias, incluyendo afroamericana, asiática, caucásica, hispana y no especificada. En este conjunto participan 48 actores masculinos y 43 femeninos. Cada actor presenta 12 frases distintas, las cuales son expresadas con las emociones siguientes: ira, disgusto, miedo, alegría, neutralidad y tristeza. Además, cada emoción se registra en cuatro niveles de intensidad diferentes: bajo, medio, alto y sin especificar [44] .

Cada archivo está identificado de manera única con un nombre con 4 pares, por ejemplo, 1001_IWW_NEU_XX.wav.

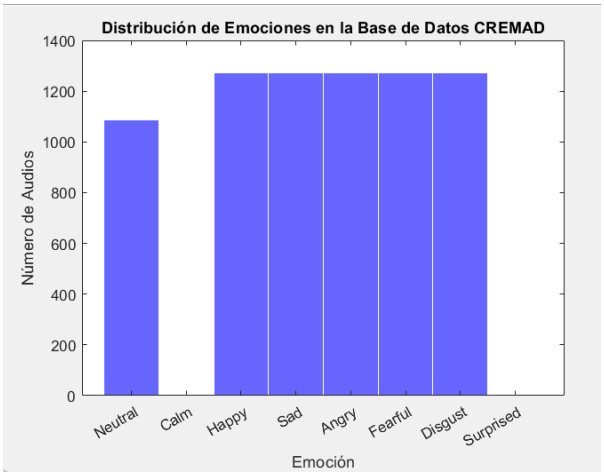
Los pares definen:

- ID del actor
- Frase que pronuncia:
    - *It is eleven o'clock (IEO).*
    - *That is exactly what happened (TIE).*
    - *I am on my way to the meeting (IOM).*
    - *I wonder what this is about (IWW).*
    - *The airplane is almost full (TAI).*
    - *Maybe tomorrow it will be cold (MTI).*
    - *I would like a new alarm clock (IWL)*
    - *I have a doctor's appointment (ITH).*
    - *Don't forget a jacket (DFA).*
    - *I think I've seen this before (ITS).*
    - *The surface is slick (TSI).*

[3] https://www.kaggle.com/datasets/ejlok1/cremad

- *We'll stop in a couple of minutes (WSI).*

- Emoción expresada: ira (ANG), disgusto (DIS), miedo (FEA), alegría (HAP), neutro (NEU), triste (SAD).
- Nivel de intensidad, el cual puede ser bajo (LO), medio (MD), alto (HI) o sin especificar (XX).




```
Número de muestras por emoción en CREMAD:
Neutral: 1087
Calm: 0
Happy: 1271
Sad: 1271
Angry: 1271
Fearful: 1271
Disgust: 1271
Surprised: 0
```

Figura 16. Visualización distribución de Emociones en la base de datos CREMA-D. (a) Gráfica de la distribución. (b) Número de muestras por emoción.

4.2. División del conjunto de datos

Independientemente de que alimentemos la red neuronal con los audios en bruto o con características, antes de introducir estos datos a la entrada de la red neuronal, interesa dividir el conjunto total de muestras creado a partir de la combinación de las bases de datos en 3 subconjuntos + 1 extra. Datos de entrenamiento, datos de validación y datos de prueba. En todos los experimentos realizados se ha mantenido una proporción constante de 70% para los datos de entrenamiento y 30% dividido entre los conjuntos de validación y prueba. El cuarto conjunto consiste en un número predefinido de muestras que la red neuronal no habrá visto durante el proceso de aprendizaje.

Los datos de entrenamiento se han utilizado para enseñar al modelo, permitiéndole reconocer patrones en las características de entrada para posteriormente asociarlos con las etiquetas de salida [45].

Por otro lado, los datos de prueba/ validación se reservan para validar el rendimiento del modelo durante el entrenamiento. Aunque estos datos se utilizan para hacer mejoras continuas en el modelo, no participan de forma directa en el proceso de aprendizaje [45]

Los datos de validación extra se mantienen apartados hasta que el modelo ha sido completamente entrenado. Su función es evaluar la eficacia del modelo en condiciones no vistas previamente, estos datos son los que proporcionan una medida más precisa de la precisión del modelo.

La elección de esta proporción 70-30 entre entrenamiento y prueba en las redes neuronales se elige cuidadosamente para lograr un equilibrio entre proporcionar suficientes datos para el aprendizaje y reservar un conjunto significativo para evaluar el modelo. Esto ayuda a evitar el sobreajuste y el infra ajuste, garantizando tamaños adecuados para ambos conjuntos. Aunque no es una regla estricta, se trata de una práctica habitual en la partición de datos para el entrenamiento y la evaluación de modelos de aprendizaje automático [46].

Hay que añadir también que, la división efectiva, se logró mediante el uso de la función '*randperm*' para generar índices aleatorios y con estos asegurar la selección aleatoria de muestras, evitando patrones sistemáticos que puedan afectar a los conjuntos.

4.3. Técnicas de procesamiento de la señal

Antes de entrenar el modelo, los datos seleccionados, se 'preparan' en distintas etapas. Este procesamiento se refiere a las acciones o técnicas realizadas sobre el conjunto antes de que se introduzcan en el modelo o antes de someterlo a un procesamiento más detallado.

Inicialmente surgió la duda sobre la aplicación de preprocesamiento a los datos extraídos de las bases de datos. Estos conjuntos de datos están diseñados para el estudio de expresiones emocionales y, en teoría, se pueden considerar

preparadas para el análisis sin necesidad de eliminar ruidos o aplicar procesamientos adicionales.

Sin embargo, a lo largo del proyecto sí se aplicaron algunas de las técnicas que se detallan en los siguientes puntos.

4.3.1. Normalización de la señal

Consiste en ajustar la amplitud de las señales de manera que todas compartan una escala uniforme. Este ajuste mantiene la consistencia en la amplitud, permitiendo así una comparación más efectiva y mejorando el rendimiento de los algoritmos.

Existen varias técnicas para normalizar los datos, y en el contexto de señales de voz, los enfoques más comunes son: la normalización *min-max*, la normalización z-score y la normalización por máximo absoluto:

La normalización *min-max* escala los valores de las amplitudes de las señales para situarlos dentro de un rango específico, normalmente entre 0 y 1, mediante esta fórmula:

$$Xnorm = \frac{\mathrm{X} - \mathrm{Xmin}}{Xma- - \mathrm{Xmin}}$$

Donde,

X es el valor original de la variable.

Xmin es el valor mínimo de la variable.

Xmax es el valor máximo de la variable.

Xnorm es el valor normalizado de la variable en el rango [0, 1].

La normalización *Z-score* o estandarización, transforma los valores de una variable para que tengan una media de 0 y una desviación estándar de 1. La fórmula utilizada es:

$$Z = \frac{\mathrm{X} - \mu}{\sigma} \tag{8}$$

Donde,

X es el valor original de la variable.

μ es la media de la variable.

σ es la desviación estándar de la variable.

Z es el valor normalizado de la variable utilizando el puntaje z.

La normalización por máximo absoluto también es muy común en el trabajo con señales de voz. Al igual que las técnicas anteriores, busca mitigar las disparidades entre las variables para garantizar una comparación efectiva. En esta técnica, los valores se escalan para que caigan dentro de un rango definido, entre -1 y 1 (mirar si debiesen estar entre 0 y 1 para luego entrenar la red porque estoy utilizando esta normalización).

La fórmula utilizada es:

$$Xnorm = \frac{\mathrm{X}}{\max(|X|)} \tag{9}$$

Donde,

x *es el valor original de la variable.*

Xnorm es el valor normalizado de la variable en el rango [-1, 1].

4.3.2. Preénfasis y Ventaneo

Ambas técnicas son comúnmente utilizadas en el procesamiento de audio. Especialmente cuando se trabaja con señales de voz, para mejorar la calidad de la información sonora.

El Filtro de Preénfasis se implementa como un filtro paso alto con el objetivo de resaltar la energía de las altas frecuencias en comparación con las bajas.

Cuando se graba o transmite una señal de audio, debido a la respuesta del micrófono y de las características acústicas del entorno, las frecuencias más altas tienden a atenuarse más que las bajas, el filtro preénfasis compensa esa atenuación realzando las frecuencias más altas y mejorando así la calidad y la eficacia del análisis.

La fórmula general para el filtro es:

$$y[n] = x[n] - \alpha \cdot x[n-1] \quad (10)$$

Donde,

y[n] es la señal de salida después de aplicar el filtro.

x[n] es la señal de entrada original.

a es el coeficiente de preénfasis, generalmente en el rango de 0.9 a 1.0.

El proceso de ventaneo igualmente se aplica sobre la señal de audio antes de aplicar técnicas de extracción de características. Consiste en dividir la señal en fragmentos superpuestos llamados '*frames*' y multiplicar cada *frame* por una función de ventana, como lo es la ventana de Hamming. Esta función de ventana atenúa los extremos del *frame* para reducir las discontinuidades en el inicio y fin del frame, reduciendo discontinuidades y mejorando la representación en el dominio del tiempo.

La fórmula para el ventaneo utilizando la ventana de *Hamming* es:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (11)$$

Además, se multiplica cada muestra por el valor correspondiente de la ventana:

$$y[n] = x[n] \cdot x[n] \quad (12)$$

Donde,

w[n] es el valor de la ventana en el instante n.

N es el tamaño de la ventana.

4.3.3. Aumento de datos

A medida que los sistemas de reconocimiento se exponen a más datos durante el entrenamiento, pueden identificar con mayor precisión las características de las entradas, lo que los hace conocidos por su enfoque 'avaricioso'. Sin embargo, uno de los desafíos principales en el aprendizaje supervisado es la necesidad de etiquetar los datos de entrenamiento para que el sistema pueda identificar su naturaleza.

Para abordar este desafío, es común utilizar la estrategia de aumento de datos de entrenamiento mediante técnicas que conservan la información relevante de las clases mientras se reduce el sobreajuste que surge cuando el sistema se ajusta demasiado a estos datos de entrenamiento y empieza a perder la capacidad de predecir casos más generales.

En la siguiente imagen se puede ver lo que sería un sobreajuste:

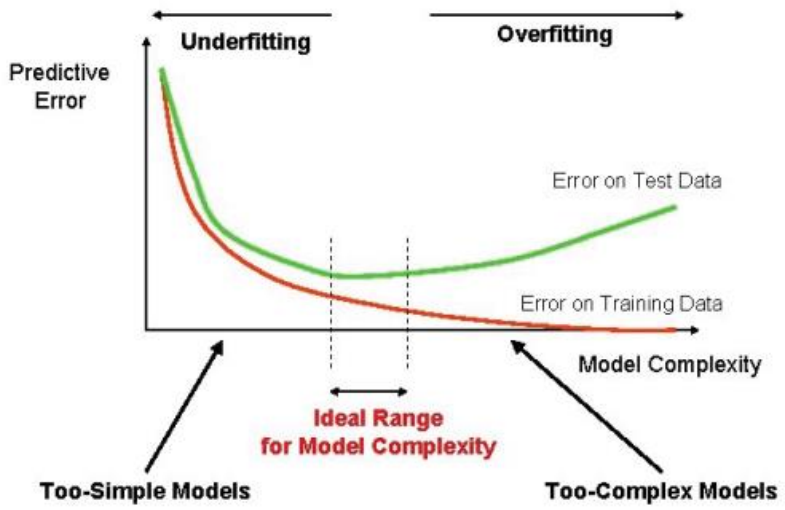


Figura 17. Representación del sobreajuste de un modelo [48]

El aumento de datos es especialmente útil cuando se dispone de una cantidad limitada de datos, ya que, al aumentar la variabilidad mediante la generación de nuevas muestras a partir de las existentes, puede mejorar la capacidad de generalización del modelo.

Estas técnicas de aumento de datos incluyen cambios en la velocidad, desplazamientos de tono y desplazamiento temporal, adición de ruido, entre otros, que representan las variaciones que pueden encontrarse en situaciones del mundo real. Al aplicar estas transformaciones de forma controlada, se construye un modelo más sólido y robusto, capaz de manejar un mayor rango de datos de entrada y evitar el sobreajuste.

4.4. Selección de características

Una vez se han procesado o no los datos, el siguiente paso es seleccionar las características más relevantes de los audios. Si bien la extracción de características no es una parte obligatoria en el proceso de detención de emociones a partir de señales de voz, normalmente este paso intermedio, mejora los resultados obtenidos.

En general, las MLPs pueden entrenarse y ejecutarse rápidamente en conjuntos de datos relativamente pequeños. Sin embargo, a medida que

aumenta el tamaño y la complejidad del problema, el tiempo de entrenamiento y la velocidad de interferencia pueden aumentar significativamente.

La extracción de características reduce la dimensionalidad del conjunto de datos, al eliminar características redundantes o irrelevantes, el modelo necesita menos datos para aprender y menos cálculos para procesar durante el entrenamiento. Al seleccionar las características más relevantes, se simplifica y agiliza el modelo. Un modelo más simple generalmente requiere menos recursos computacionales y puede hacer predicciones más rápidas.

Para comprender cómo funcionan estos métodos, primero es necesario explicar la base matemática en la que se sustentan, que es la Transformada de Fourier (FT):

Una señal de audio está formada por diversas ondas sonoras, cada una con su propia frecuencia. La FT es una operación matemática que descompone una señal en sus frecuencias individuales, convierte la señal del domino del tiempo al dominio de la frecuencia mediante el uso de senos y cosenos. La salida de esta operación se conoce como espectro.

$$X(f) = \int_{-\infty}^{\infty} x(t) \cdot e^{-j2\pi ft}\ dt \qquad (13)$$

La Transformada rápida de Fourier (FFT) es una versión más eficiente de la FT, igualmente descompone la señal en sus componentes de frecuencia, pero se utiliza para analizar la frecuencia de una señal de audio de manera más rápida, siempre y cuando, las señales sean periódicas y no cambien con el tiempo.

La técnica de ventaneo desarrollada en un apartado anterior deja unos segmentos a los que se puede aplicar la FFT para obtener un espectro de frecuencia para cada uno de ellos. Al superponer estas ventanas y combinar los espectros resultantes, se obtiene lo que se conoce como espectrograma, una representación visual que muestra cómo varía la energía de las diferentes frecuencias a lo largo del tiempo.

También es importante mencionar la Transformada de Fourier de Tiempo Corto (STFT) que, en lugar de analizar toda la señal en su conjunto, divide la señal

en segmentos más pequeños y aplica sobre cada uno la FT, proporcionando una representación de cómo varían las características de frecuencia a lo largo del tiempo [49].

Una vez que se tienen estas transformadas y el espectrograma, se pueden extraer diferentes características relevantes para el reconocimiento. Se presentan a continuación algunas de las más comunes:

- **Frecuencia fundamental F0 (*pitch*):**

La frecuencia fundamental se refiere a la vibración de las cuerdas vocales y determina el tono o la altura de la voz. No solo incluye la representación de la frecuencia en sí misma, también el contorno de esta describe las variaciones del tono en forma de patrones geométricos.

Si se quisiera representar una gráfica que muestre cómo cambia la frecuencia fundamental de una persona mientras habla, se observaría que no es una línea recta, sino que presenta subidas y bajadas que representan las variaciones del tono de voz. Por ejemplo, al hacer una pregunta, que se da una entonación al final de la oración, en la gráfica se vería como un aumento de la frecuencia.

- ***Chroma*:**

Se trata de una representación tonal y armónica de una señal de audio que captura información sobre la distribución de tonos musicales en diferentes frecuencias o a lo largo del tiempo. Para calcular esta característica se aplica la STFT sobre la señal de audio, lo que permite analizar cómo cambian las características tonales en función del tiempo y la frecuencia.

- **Espectrograma de Mel:**

Se trata de otra representación del espectro de frecuencia. Las personas no perciben las frecuencias en una escala lineal, sino que son más sensibles a los cambios en frecuencias más bajas que en las más altas. Para abordar esta

diferencia en la percepción auditiva, se ideó en 1937 la escala de Mel. Esta escala proporciona una medida tonal donde las diferencias de tono se perciben de manera uniforme para el oyente.

Esta herramienta transforma una frecuencia dada a una medida en la escala de Mel. El hecho de que permita imitar la manera en que los seres humanos perciben el sonido hace que se convierta indispensable en aplicaciones de aprendizaje automático, como el reconocimiento de voz o emociones.

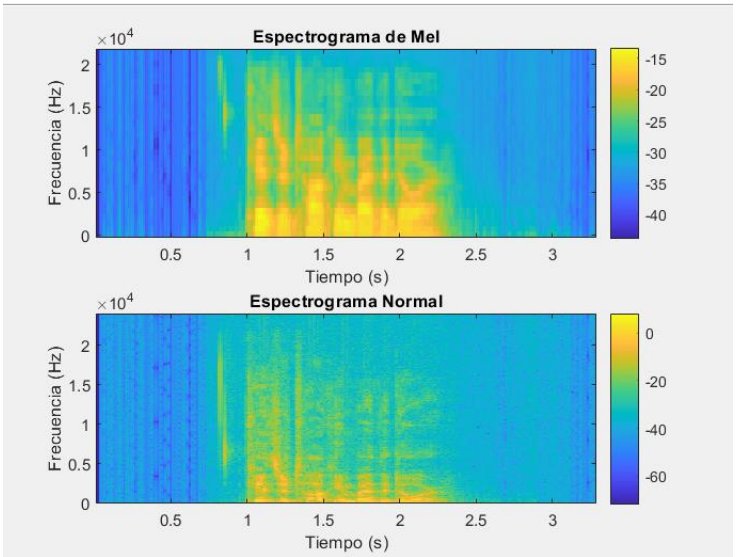


Figura 18. Comparativa entre la escala de Mel y la normal de un audio

- **MFCC:**

Los Coeficientes Cepstrales de Mel se derivan de la STFT y están diseñados para capturar la información relevante de la señal de voz y representar de manera eficaz las características perceptuales del habla según la percepción humana.

El propósito de estos es extraer características relevantes de las componentes de una señal de audio, poniendo el foco en la información más significativa y desestimando contenido menos validos como ruidos de fondo.

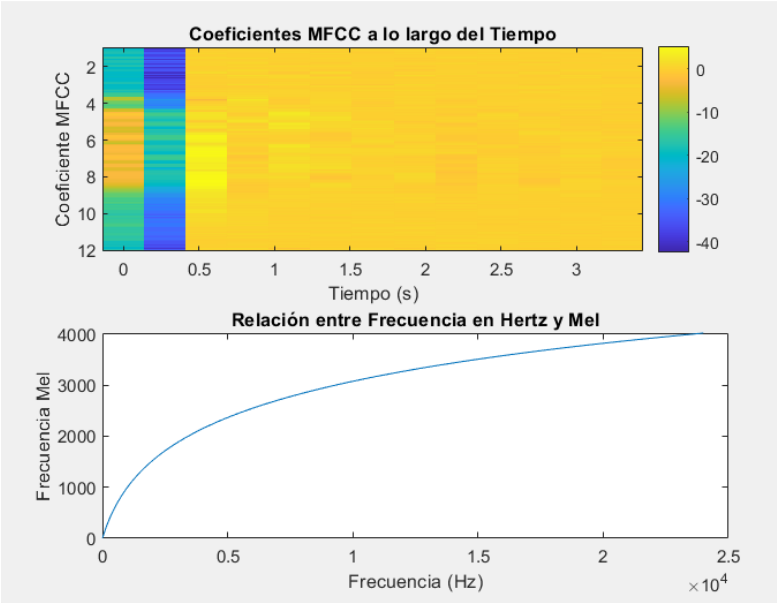


Figura 19. Representación MFCC de uno de los audios del conjunto de datos en Matlab.

- **ZCR:**

La Tasa de Cruces por Cero es una medida que representa la cantidad de veces que una señal cambia de polaridad, es decir, de positiva a negativa o viceversa. Un ZCR mayor indica que la señal tiene más cambios de dirección, lo que puede ser por la presencia de componentes de alta frecuencia o ruido. Para calcularlo, se promedian los valores a lo largo del eje 0 de la señal.

- **RMS *(Root Mean Square)*:**

El valor eficaz RMS cuantifica el nivel promedio de amplitud de una señal. Se calcula tomando la raíz cuadrada de la media de los cuadrados de las amplitudes de la señal. En otras palabras, el RMS nos proporciona una medida de la 'potencia' promedio de la señal, considerando tanto valores positivos como negativos y proporcionando una medida que representa la energía total de la señal de forma más equilibrada.

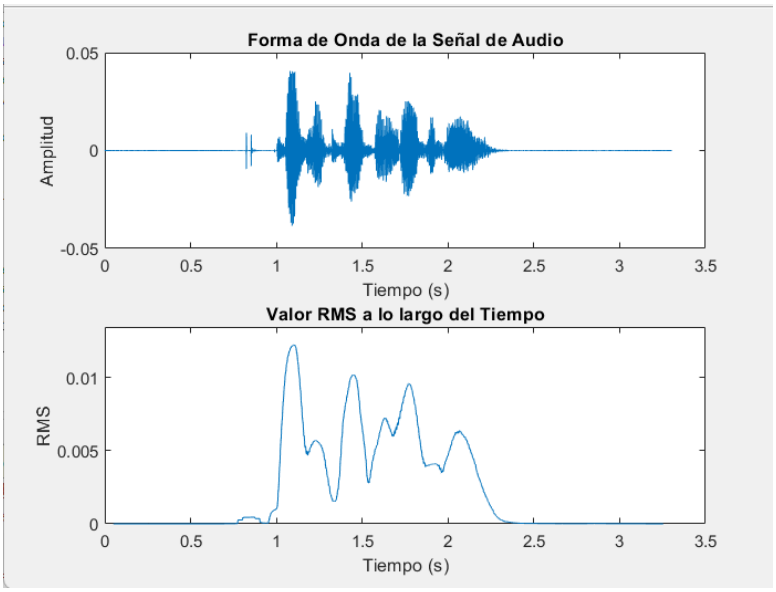


Figura 20. Representación RMS de uno de los audios del conjunto de datos en Matlab.

- **Intensidad de la voz:**

Se relaciona con cuanto de fuerte o débil es el sonido que se está produciendo. Se mide en decibelios (dB) y proporciona información sobre el nivel de energía acústica presente en la señal de voz. Una mayor intensidad indica que la voz es más fuerte y una intensidad más baja indica una voz más suave y con menos energía.

Hay emociones como la ira o la euforia que asociamos como emociones intensas, debido a un aumento en el nivel de energía de la señal de voz. Emociones más suaves como la tristeza o la calma, reflejan una intensidad vocal más baja.

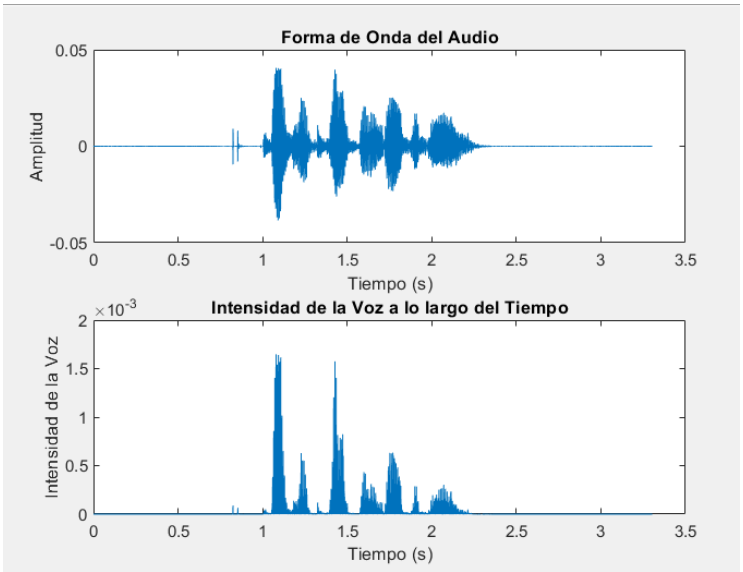


Figura 21. Representación intensidad de uno de los audios del conjunto de datos en Matlab.

- **Shimmer:**

La característica *Shimmer* es una medida que se utiliza en señales de voz para cuantificar la variabilidad en la amplitud de las ondas sonoras, es decir, para medir cómo cambia la altura, la amplitud en cada ciclo vocal.

Para calcular el *Shimmer* de amplitud, se mira cada ciclo vocal y se mide la diferencia entre la parte más alta y la más baja de sonido (amplitud pico a pico). Luego, se toman las diferencias de todos los ciclos y se suman y dividen entre la cantidad de ciclos para obtener un promedio.

Un *Shimmer* alto significará mucha variabilidad, lo que indica que la amplitud cambia mucho.

- **Formantes:**

Como ya se mencionó en el estudio teórico, los formantes son resonancias que se producen en el tracto vocal cuando se emite una señal de voz y que dependen por lo tanto de la forma y dimensiones de este. Es una característica que se refleja en el dominio de la frecuencia, cada formante está caracterizado por su frecuencia central y su ancho de banda.

Para estimar estos dos parámetros, es común utilizar un análisis de predicción lineal. Se trata de una técnica donde se modela el tracto vocal mediante un

polinomio de orden M que contiene polos que representan las resonancias del tracto vocal. Los coeficientes de predicción lineal (LPC) se utilizan para determinar la ubicación de los polos, es decir, la frecuencia central de cada formante y su ancho de banda en cada ventana de la señal de voz.

- **Operador de energía Teager o Cuadrática:**

Esta característica se utiliza para detectar cambios rápidos, bruscos o impulsos en la señal de voz. Se calcula tomando la señal de voz y aplicando una operación no lineal en el dominio del tiempo para resaltar las componentes de energía de alta frecuencia.

Para calcular el operador de energía de *Teager* se siguen los siguientes pasos:

1. Se toma un valor de la señal x(n) en un instante concreto.
2. Se calcula el cuadrado de este valor x(n)^2.
3. Se multiplican los valores de la señal en 2 puntos adyacentes, uno en el instante anterior x(n-1) y otro en el instante siguiente x(n+1).
4. Se multiplica todo lo mencionado anteriormente.

$$Et(n) = x(n)^2 - x(n-1) \times x(n+1) \quad (14)$$

- **Energía de la señal de voz:**

La energía en una señal de voz se refiere a la amplitud o intensidad contenida en la señal en diferentes intervalos de tiempo.

Para calcular esta se suma cada muestra de la señal de voz y luego se toma el promedio de estos valores a lo largo de una ventana de tiempo. Matemáticamente, para una señal de voz x(n) y una ventana de tamaño N, la energía E(n) se calcula como:

$$E(n) = \frac{1}{N} \sum_{i=n}^{n+N-1} x(i)^2 \quad (15)$$

Ciertas emociones como la felicidad o la ira se expresan con mayor energía vocal, mientras que la tristeza o el cansancio reflejarán un valor de energía más bajo, por eso es interesante estudiar este parámetro en el contexto del reconocimiento de emociones.

- **LFPC (Log Frequency speech power Coefficients):**

Los coeficientes Cepstrales de frecuencia lineal describen la forma del espectro de la señal. Antes de calcularlos se ha de enventanar cada segmento de la señal para aplicar a cada uno un peso distinto además de reducir el ensanchamiento del espectro.

Se lleva al dominio de la frecuencia mediante la STFT y se utilizan generalmente, 12 filtros paso banda sobre sus componentes. Después, se obtienen las energías de cada uno de los filtros y para reducir su amplitud y mejorar la representación de la información, se toma el logaritmo de cada energía calculada.

Finalmente, se aplica la Transformada Cepstral inversa para obtener los coeficientes, es una transformación matemática que convierte el espectro de la frecuencia en el dominio cepstral, más adecuado para el procesamiento de voz.

4.5. Detalles de la implementación

A continuación, previo a desarrollar el proyecto, se muestra el marco iterativo que me ha permitido ajustar y adaptar continuamente el modelo a medida que avanzaba en mi investigación.

Mi enfoque se centró en desarrollar un modelo capaz de distinguir entre ocho emociones utilizando la base de dato RAVDESS, con el fin de enriquecer este conjunto de datos y de mejorar el rendimiento del modelo, se decide agregar dos bases de datos adicionales, SAVEE y CREMA-D.

Todos los códigos tienen en común la fase inicial, donde se genera una matriz de etiquetas y otra matriz de datos:

En primer lugar, se obtienen los nombres de los archivos de audio de la base de datos RAVDESS y a partir de esta se crean 2 nuevas matrices de etiquetas, una de ellas en formato numérico simple y otra en formato '*one-hot encoding*'. En este último formato, cada emoción se representa con un vector binario, donde solo uno de los elementos es 1 y los demás son 0. Utilizando esta codificación, se evita que el modelo asuma que hay un orden de importancia entre las clases. El modelo no interpretará incorrectamente que una emoción con un número alto es 'más importante' que otra con un número más bajo. Además, después de hacer varias pruebas, mi modelo de red neuronal trabaja mejor con etiquetas en este formato.

Al utilizar *one-hot encoding*, cada emoción está representada como un vector binario de longitud igual al número total de clases (8 por ejemplo). Cuando la red realiza la predicción, la salida de la capa de salida también será un vector de 8 elementos, donde cada elemento representa la probabilidad de pertenecer a esa emoción.

Con esta codificación se garantiza la compatibilidad asegurando que la estructura de la salida de la red y las etiquetas de entrenamiento coincidan. Este enfoque ha demostrados ser muy útil para la clasificación multiclase.

Por ejemplo, la emoción '*happy*' (03-01-03-01-02-01-12.mp4) asociada al valor 3 en el tercer par del nombre del audio, estaría representada como [0 0 1 0 0 0 0 0] en la codificación *one-hot*.

Lo mismo con las otras 2 bases de datos para crear una matriz de etiquetas combinada.

La segunda fase, también común en todos los entrenamientos ha sido encontrar el tamaño mínimo entre todos los audios de cada base de datos y ajustar los demás a este. También se contempló la opción de coger el tamaño promedio y ajustar rellenando o eliminando en función del audio, pero los modelos no presentaban mejora.

Una vez ajustados todos los audios al mismo tamaño en cada base de datos, se plantearon dos caminos para construir la matriz de datos combinada, en algunas pruebas se encontró la longitud mínima entre las 3 matrices de audios para ajustar de nuevo todos los audios al tamaño mínimo; y en otras pruebas se utilizó un tamaño promedio y en función de este, se ajustaron el resto de los audios mediante el relleno o el recorte correspondiente.

A continuación, se examinan detalladamente las partes que han experimentado variaciones a lo largo del proyecto, y que sirven como base para el desarrollo de todos los experimentos.

Se parte de una RNA, con la siguiente arquitectura [Figura 22]:

1. Capa de entrada: No se especifica la cantidad de muestras en la capa de entrada, se parte de una matriz de audios en bruto de tamaño 9362 x 20287
2. Capas intermedias: 2 capas de 50 neuronas con función de activación '*Softmax*'.
3. Capa de salida: 8 neuronas que se corresponden con las 8 emociones que se quieren clasificar.

Figura 22. Diagrama de la arquitectura de la red neuronal inicial con 2 capas ocultas (experimento P1)

Y con los siguientes parámetros:

- **net.trainParam.epochs = 6000**. Especifica la cantidad máxima de veces que la red neuronal recorrerá todo el conjunto de entrenamiento. Una época completa significa que la red ha visto y se ha entrenado en todos los ejemplos del conjunto de datos.

- Se establece un objetivo de rendimiento deseado **net1.trainParam.goal = 0**; lo que significa que la red se entrenará durante el número especificado de épocas sin considerar un objetivo específico de rendimiento, inicialmente se puso así para ver cómo iba funcionando el sistema.

- En la compilación del modelo se utiliza una tasa de aprendizaje del optimizador Adam **net.trainParam.lr = 0.001**, este valor representa la tasa de aprendizaje en el contexto del entrenamiento de la red. se trata de un algoritmo de optimización que controla 'cuanto aprende' la red en cada iteración del proceso de entrenamiento. Si la tasa es demasiado pequeña, el entrenamiento pude ser lento, y si es demasiado grande, el entrenamiento puede volverse inestable.

Entonces, aunque se establezca ese valor para la tasa de aprendizaje como punto de partida, el optimizador Adam ajustará esta tasa de aprendizaje durante el proceson1 según las estadísticas de los gradientes que observa y lo adaptará para mejorar la convergencia y eficacia del aprendizaje.

- Con **net.trainParam.min_grad** se establece como estrategia un umbral para el gradiente. Si el gradiente de los pesos de la red cae por debajo de este umbral, el entrenamiento puede detenerse, ya que podría indicar que la red ha convergido o que está avanzando muy lentamente.

- El parámetro **net.trainParam.max_fail**. Si el rendimiento en el conjunto de validación no mejora durante un número consecutivo de intentos, el entramiento se detiene para evitar sobreajuste.

Todas las redes neuronales desarrolladas en este proyecto utilizan la función '*train*' para el entrenamiento de los datos de entrada, ya sean los audios en bruto o la matriz de características, y sus etiquetas correspondientes.

En MATLAB, esta función está diseñada para ser utilizada con distintos tipos de redes neuronales, se trata de una herramienta versátil y compatible con diferentes arquitecturas de redes.

A partir de este primer modelo se desarrollaron las siguientes arquitecturas para entrenar las diferentes redes neuronales:

Figura 23. Diagrama de la arquitectura de la red neuronal (experimento P2), con audios en bruto como entrada

.

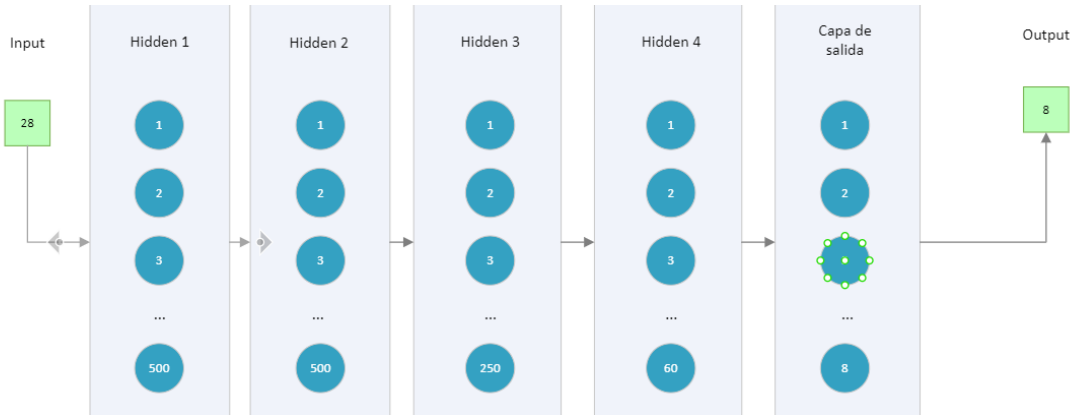


Figura 24. Diagrama de la arquitectura de la red neuronal (experimentos P3-P6), con matriz de características como entrada.

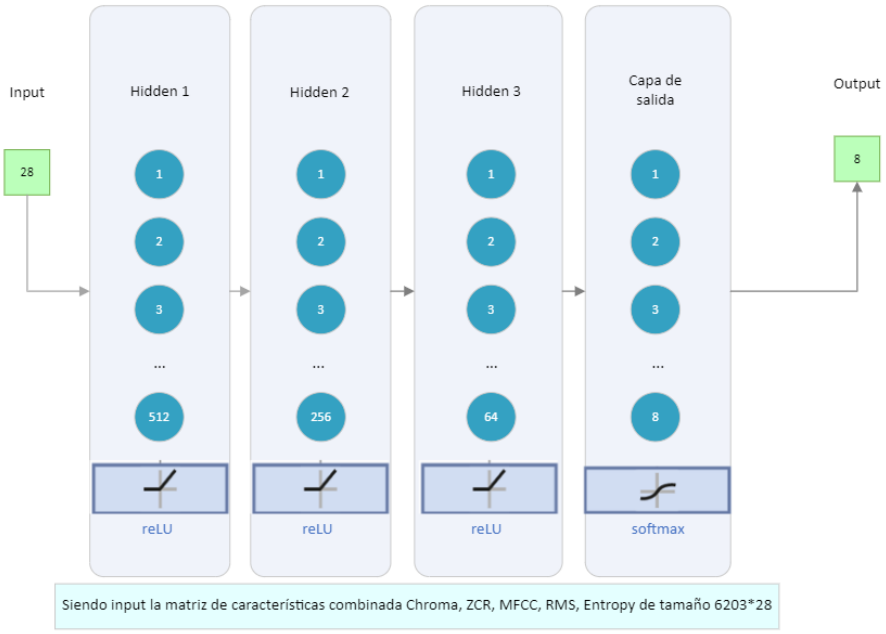


Figura 25. Diagrama de la arquitectura de la red neuronal (experimento P7).

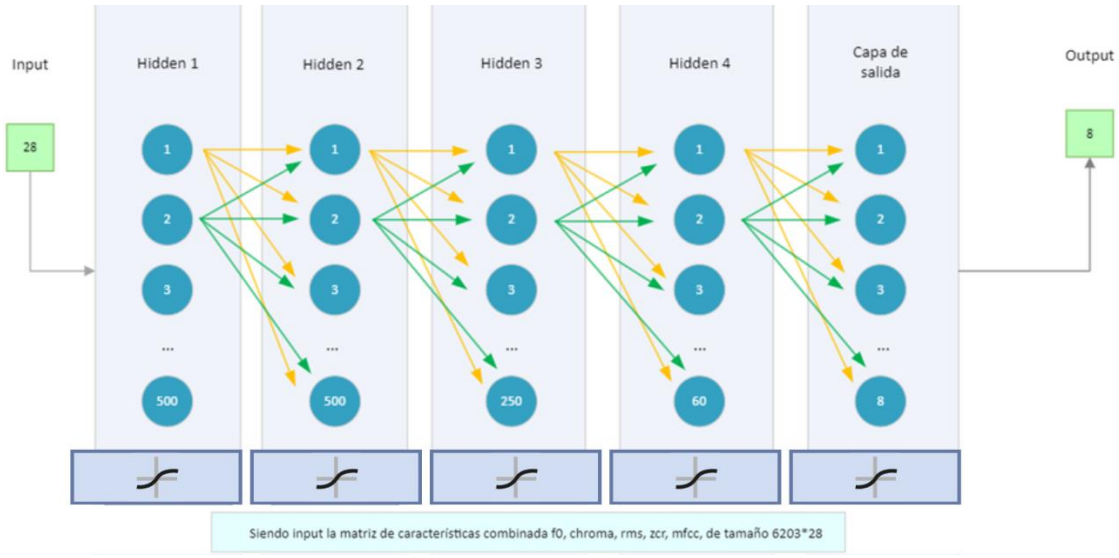


Figura 26. Diagramas Experimento con los mejores resultados P6z.Diagrama detallado feedforward de la arquitectura de la red neuronal.

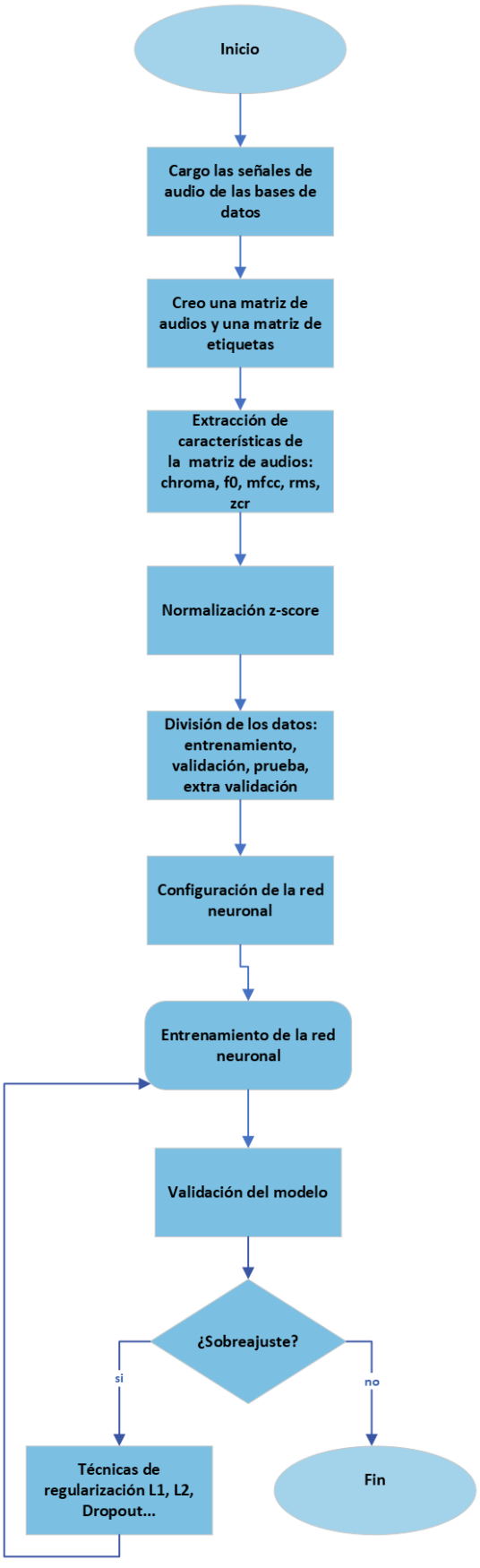


Figura 27. Diagramas Experimento con los mejores resultados P6z..Flujo creación de la red neuronal

5. RESULTADOS

En este apartado de la memoria se detallan los resultados obtenidos en cada modalidad utilizando diferentes estructuras y enfoques. Se ha seguido una proporción común de datos de entrenamiento, validación y prueba del 70%, 15% y 15% respectivamente, dejando x número de muestras apartadas para posteriormente validar el modelo, como se explicó en la sección 4.2 del documento.

Se modificó progresivamente el tamaño del conjunto de muestras incorporando dos bases de datos adicionales para aumentar la cantidad de audios. No se aplica la técnica de aumento de datos debido a que el modelo no mostraba suficiente robustez como para agregar nuevos audios.

Durante el proceso de experimentación, se realizaron modificaciones en parámetros clave, como la frecuencia de muestreo (fs) y el número de características seleccionadas. Se experimentó también con diferentes enfoques de entrada, incluyendo la utilización de audios en bruto y características extraídas.

Para mejorar el desempeño de los modelos, se variaron las arquitecturas de las redes neuronales y se introdujeron técnicas de normalización y regularización.

Todos estos experimentos se han documentado en una hoja de cálculo Excel para su análisis detallado en una sección posterior del trabajo.

CLASIFICADOR DE 8 EMOCIONES

Las emociones por reconocer son: sorpresa, ira, calma, disgusto, tristeza, miedo, alegría y neutral. Inicialmente se utiliza la base de datos RAVDESS, pero progresivamente se van añadiendo también las bases CREMA-d y SAVEE, consiguiendo un total de 9362 archivos .wav

Una vez creadas las matrices combinadas de datos y de etiquetas se decide plantear dos enfoques, por un lado, se decide entrenar el modelo con

los audios en bruto y, por otro lado, se opta por utilizar las características extraídas de estos. Los audios en bruto conservan toda la información acústica de la muestra, lo que es beneficioso si hay patrones complejos, sin embargo, tiene un mayor uso computacional. Por otra parte, al entrenar el modelo con características, se reduce la dimensionalidad del conjunto de datos, acelerando el entrenamiento, las características pueden proporcionar una interpretación más clara, pero se pierde parte de la información original.

Esta adaptabilidad ha sido clave, ya que algunos experimentos demostraron mejor rendimiento con audios en bruto, mientras que otros mejoraron al utilizar las características extraídas. LA elección entre ambos métodos se guio por la experimentación continua, gracias a la flexibilidad de este enfoque fueron mejorando los resultados según las necesidades específicas de cada banco de pruebas.

En el Excel de pruebas se han documentado todas las configuraciones, resumen, resultados y observaciones de más de 20 experimentos. En este apartado se intenta detallar lo que se ha observado durante el proyecto, desarrollando más a fondo las técnicas empleadas y los porcentajes de precisión obtenidos en los experimentos con mejores resultados.

Los primeros clasificadores de ocho emociones se entrenaron con los audios en bruto extraídos de las 3 bases de datos mencionadas anteriormente. Sin ningún tipo de procesamiento previo, los datos se dividieron en los conjuntos de entrenamiento, validación y prueba, además del conjunto adicional para la validación extra (explicado en el punto 4.2).

Para la implementación, se partió de la red neuronal inicial mencionada un par de apartados más arriba. La capa de salida se configuró con ocho neuronas correspondientes a las ocho emociones estipuladas como objetivo en la clasificación multiclase.

Se utilizó la función *patternet* para crear la red neuronal de retropropagación y la función *train* para entrenar la red con los datos de entrenamiento seleccionados.

P1:

Se utiliza la arquitectura base mencionada en los detalles de la implementación, se probaron diferentes funciones de activación hasta que se logró una precisión cercana al 18% en el conjunto de evaluación y un 85.10% en los datos de entrenamiento utilizando la función de activación '*softmax*' para la capa de salida. Estos resultados fueron muy interesantes, ya que aún no se había aplicado ningún tipo de procesamiento adicional a los datos ni se habían utilizado técnicas para controlar el entrenamiento.

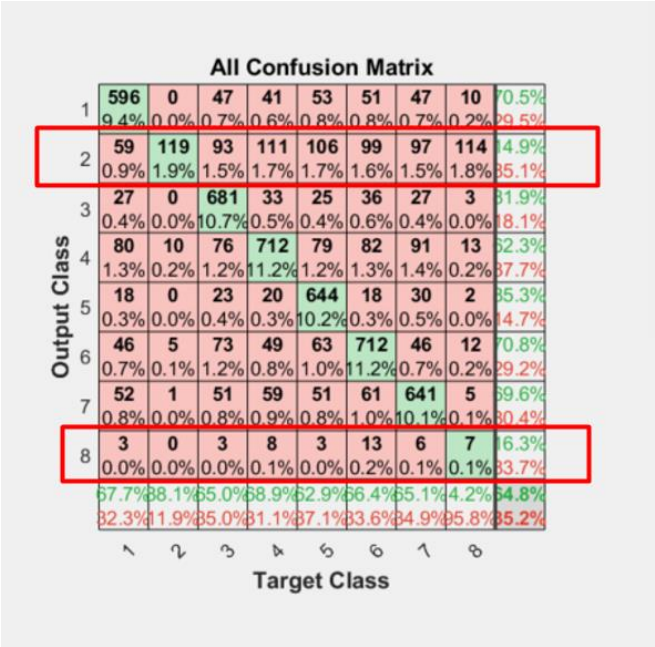


Figura 28. All Confusion Matrix Experimento P1

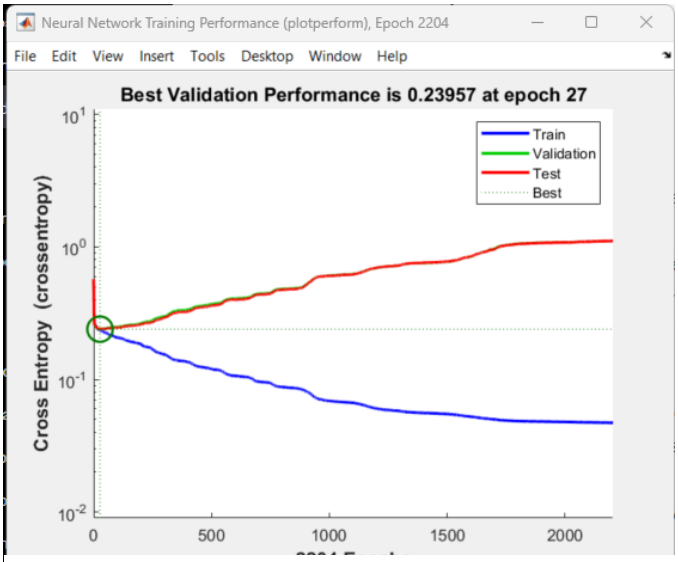


Figura 29. Performance Experimento P1.

Observando las gráficas de este primer experimento recogido también en el Excel de pruebas, se identificaron dos aspectos que podían estar perjudicando el entrenamiento.

En primer lugar, la gráfica del rendimiento que ofrece MATLAB revelaba un evidente sobreajuste, mientras que la matriz de confusión sobre todo el conjunto de datos indicaba un posible desequilibrio entre las clases.

Para hacer frente a estos problemas se plantearon 3 cuestiones. En primer lugar, con el objetivo de mejorar el bajo rendimiento especialmente de las clases 2 y 8 en el conjunto de validación y contrarrestar el sobre ajuste, se exploraron técnicas como regularización, *Dropout*, extracción de características….

Además, se propuso realizar un análisis detallado de las clases 2 y 8 para encontrar qué podía estar causando dificultades en su clasificación.

Por último, se planteó la experimentación con diferentes arquitecturas de red y parámetros de entrenamiento para encontrar la combinación óptima que mejorase la precisión global.

En el siguiente banco de pruebas **P2**, se añaden 2 capas ocultas, resultando en un modelo de 500, 500, 250, 60 neuronas por capa. Además, se configura una función de activación *'logsis'* para las capas ocultas y se mantiene la función *'softmax'* en la capa de salida.

Para hacer frente al sobreajuste del modelo anterior, se aplicó la *regularización L2(Ridge)* manualmente mediante el parámetro Lambda, puesto que, en Matlab, la función *patternet* no admite directamente la regularización L2. Esta técnica agrega un término adicional a las capas densas del modelo, se conoce como término de peso de decaimiento o regularización de Ridge. Se emplea con la finalidad de 'castigar' los pesos más grandes evitando que ciertos pesos dominen la red y se ajusten excesivamente a los datos de entrenamiento.

Sin embargo, esta técnica no mejoró los resultados y también se podía percibir un sobreajuste. Durante las 4 primeras horas de entrenamiento, la precisión en el conjunto de entrenamiento no alcanzaba el 52%, y la precisión en el conjunto de validación rondaba el 15%, después de 12h de entrenamiento, la precisión disminuyó hasta el 21.75% para los datos de entrenamiento y hasta el 16% en el caso de los datos de validación.

Viendo que lejos de mejorar, los resultados empeoraban, se decidió tomar otro camino. Por una parte, se analizó detalladamente el problema con el reconocimiento de las clases 2 y 8, correspondientes a las emociones 'calma' y 'sorpresa' respectivamente y se atribuyó al desbalanceamiento inherente del conjunto de datos. Este desequilibrio se origina porque las bases de datos Crema D y SAVEE no incluyen muestras que representen estos estados emocionales. La baja cantidad de muestras en comparación con las otras 6 emociones afectaba negativamente a la capacidad del modelo para reconocer estas clases.

```
Distribución de clases antes de la división:
        1303         192        1523        1523        1523        1523        1523         252
```

Figura 30. Distribución de cada emoción en el conjunto total de datos

```
Distribución de clases en el conjunto de entrenamiento:
         861         135         975        1019        1011        1039        1003         160

Distribución de clases en el conjunto de validación:
   191    26   226   200   219   206   215    46

Distribución de clases en el conjunto de prueba:
   177    23   238   215   214   209   213    40
```

Figura 31. Distribución de cada emoción en los conjuntos de entrenamiento, validación y prueba.

Por otra parte, se tomó la decisión de alimentar la red neuronal con características en lugar de utilizar las muestras de audio en bruto. Esta elección se basó en primer lugar en la premisa de que proporcionar a la red información más específica y relevante podría mejorar su capacidad para discernir

patrones; otra motivación fue la reducción del tiempo de entrenamiento. Al utilizar características en lugar de muestras en bruto, se redujo la dimensionalidad de los datos de entrada, acelerando significativamente el proceso de aprendizaje.

A continuación, se detallan las técnicas empleadas a lo largo de la investigación en cada experimento:

De aquí en adelante la atención se centra en la extracción de características, en el experimento **P3** se incluyeron: cromagrama, frecuencia fundamental (F0), tasa de cruces por cero (ZCR), coeficientes cepstrales de frecuencia mel (MFCC), y RMS (*Root Mean Square*). Además, se aplicó una ventana de Hamming con un tamaño 2048 muestras, un tamaño de paso (*HopSize*) de 512 muestras y se introdujo el preénfasis de primer orden para realzar las frecuencias más altas.

```
% EXTRACCIÓN DE CARACTERÍSTICAS con los siguientes parámetros globales

windowSize = 2048; % Tamaño de la ventana en muestras
hopSize = 512; % Tamaño del paso en muestras
% Ventaneo de la señal (hsmming)
window = hamming(windowSize, 'periodic');
```

Figura 32. Ventaneo haming de la señal.

```
% Preénfasis de la señal
preEmphasized = filter([1 -0.97], 1, audio);
```

Figura 33. Filtro preénfasis.

La función de activación para las capas ocultas no se especifica, siendo la función predeterminada en MATLAB para la red patternet la función sigmoide *'logsig'*. Además, no se emplean técnicas de normalización ni regularización en este contexto.

El entrenamiento de la red se realiza con el método de retropropagación utilizando el entrenador '*trainlm*' (entrenador de Levenberg-Marquardt). Este algoritmo tiene el objetivo de minimizar la función de costo, que cuantifica la discrepancia entre las salidas producidas por la red neuronal y las salidas deseadas para el conjunto de entrenamiento. En otras palabras, su función principal es encontrar la configuración óptima de pesos que minimiza la diferencia entre las predicciones de la red y los valores reales esperados.

En **P4** se retorna a la regularización L2, se mantuvo la misma combinación de características, pero este experimento incorpora la *normalización z-score* a las características extraídas y, por otra parte, se reduce el número máximo de épocas y de fallos consecutivos para que no se produzca el sobreentrenamiento que puede verse en la Figura 23.

Después de casi 16 horas de entrenamiento se decide parar viendo que los resultados no estaban mejorando significativamente:

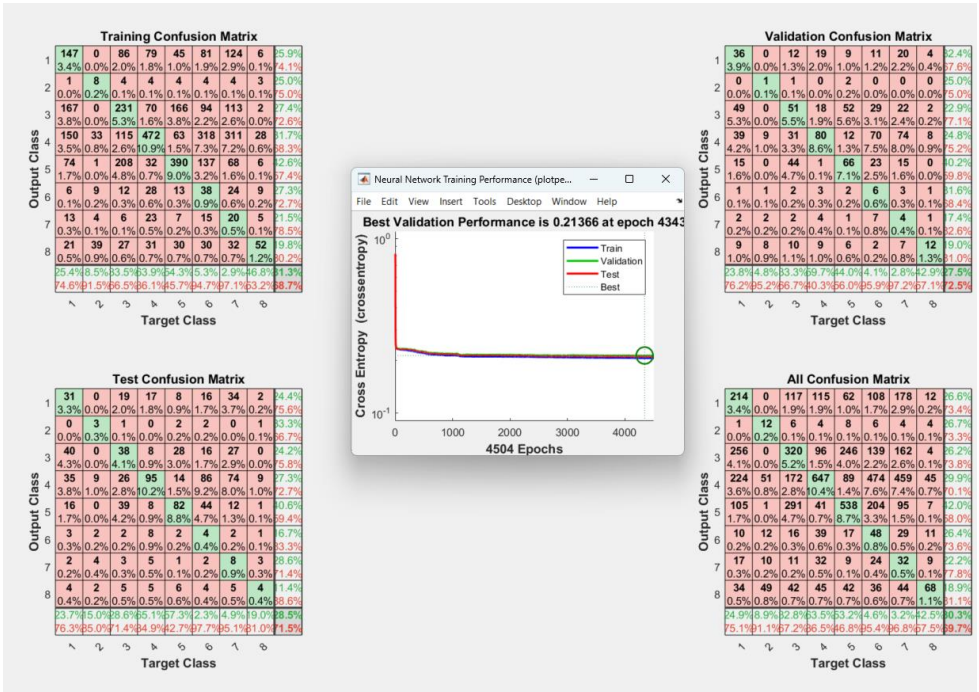


Figura 34. Performance y matriz de confusión experimento P4.

```
>>
Porcentaje de emociones correctas detectadas en el conjunto de validación: 28.4424%
Porcentaje de emociones correctas detectadas en el conjunto extra de validación: 29.6%
Porcentaje de emociones correctas detectadas en el conjunto de entrenamiento: 30.2918%
```

Figura 35.Resultados experimento P4

Como se puede ver en los resultados, esta técnica ha evitado el sobreajuste, pero no ha mejorado la precisión del modelo.

P5 utiliza el mismo código que P4, pero sustituye la regularización L2 por la regularización L1.

Ambas técnicas se utilizan para prevenir el sobreajuste introduciendo términos adicionales en la función de costo que la red está optimizando durante el entrenamiento.

Sin embargo, L1 es común si se sospecha que solo algunas características son realmente importantes para el reconocimiento y se quiere 'apagar' las menos importantes.

Si se cree que todas las características contribuyen de alguna manera al reconocimiento, L2 mantiene todos los pesos pequeños y rara vez los 'apaga' completamente.

De nuevo cumplimos el objetivo de evitar el sobre ajuste, pero la red empeora la precisión no alcanzando los resultados obtenidos en el experimento anterior utilizando L2.

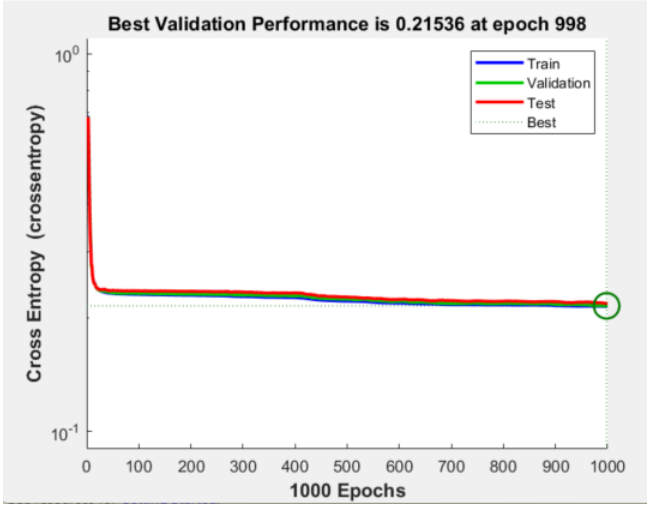


Figura 36. Performance experimento P5

El modelo **P6** mantiene la red neuronal de 4 capas ocultas con [500, 500, 250, 60] neuronas respectivamente, y una capa de salida con 8 emociones

correspondientes a las clases emocionales a detectar. Durante el entrenamiento, se aplica la regularización manual L2 puesto que ha sido la técnica que mejores resultados ha dado.

Se utiliza preprocesamiento con *one-hot encoding* y normalización z-score y se establece un entrenamiento de 6000 épocas y una tasa de aprendizaje de 0.001, al evaluar la precisión en los diferentes conjuntos se obtuvieron los resultados que se detallan a continuación.

```
%% Entrenamiento y ajuste de parámetros de la red neuronal

% Creo la red neuronal patternnet
hiddenLayerSizes = [500, 500, 250, 60];
net6 = patternnet(hiddenLayerSizes);

net6.layers{end}.size = 8; % Tamaño de la salida

% ajustamos manualmente la regularización L2
lambda = 0.01; % Parámetro de regularización L2
net6.performParam.regularization = lambda;


% parámetros de entrenamiento
net6.trainParam.epochs = 6000;  % Número de épocas
net6.trainParam.goal = 0;  % Objetivo de rendimiento
net6.trainParam.lr = 0.001;  % Tasa de aprendizaje
net6.trainParam.min_grad = 1e-6; % Criterio de convergencia basado en el gradiente
net6.trainParam.max_fail = 1000000000;  % Número máximo de fallos consecutivos

% Entreno la red neuronal
net6 = train(net6, trainingFeatures', trainingLabelsFeatures');

% Guardo la red
save('red_neuronal_entrenada.mat', 'net6');
```

Figura 37. Arquitectura y entrenamiento de la red experimento P6.

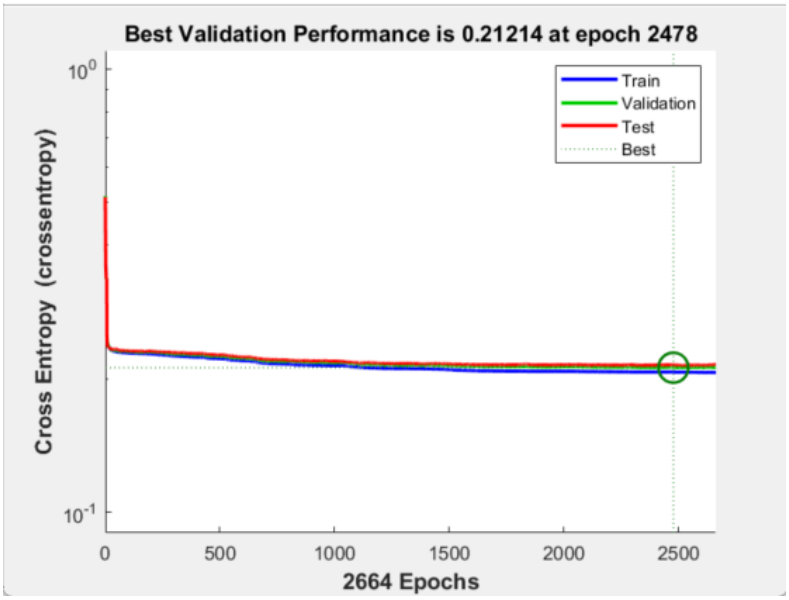


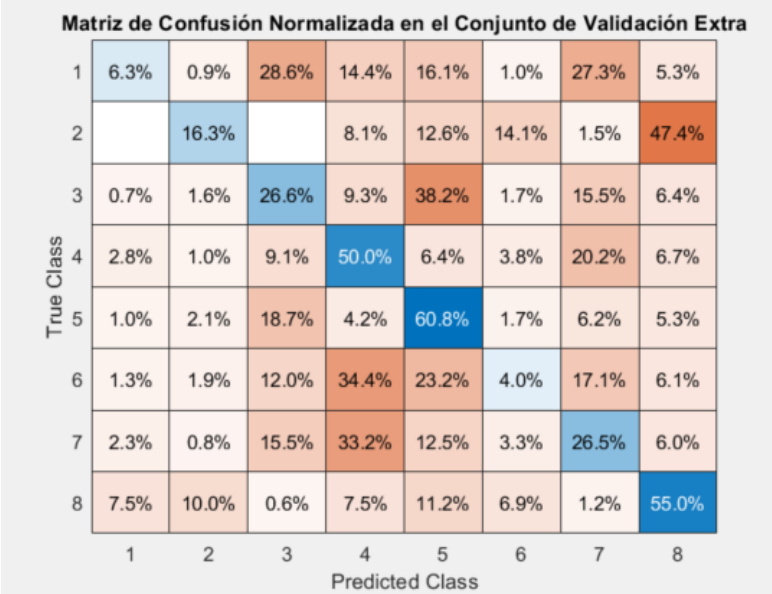
Figura 38. Performance experimento P6.

Estos fueron los resultados obtenidos:

```
Porcentaje de emociones correctas detectadas en el conjunto de entrenamiento: 29.9049%
Porcentaje de emociones correctas detectadas en el conjunto extra de validación: 30.2%
Porcentaje de emociones correctas detectadas en el conjunto de validación: 27.6147%
```

Figura 39. Resultados experimento P6.

A partir de la precisión sobre el conjunto extra de datos de validación se crea su matriz de confusión correspondiente:




```
Matriz de Confusión en el conjunto extra de validación:
    54     8   246   124   139     9   235    46
     0    22     0    11    17    19     2    64
     7    16   259    91   372    17   151    62
    29    10    93   509    65    39   206    68
    10    21   189    42   615    17    63    54
    13    20   125   357   241    42   178    63
    23     8   155   333   125    33   266    60
    12    16     1    12    18    11     2    88
```

Figura 40. Matrices de confusión experimento P6. (a) matriz de confusión normalizada y (b) matriz de confusión normal.

La red neuronal demostró tener éxito al clasificar algunas instancias de manera precisa, pero también presenta errores significativos. Las clases 3,4,5 y 7 obtuvieron el menor número de errores de clasificación, alcanzando hasta un 60% de predicciones correctas. Este rendimiento sugiere que la Red está logrando clasificar estas clases con una exactitud relativamente alta.

Por otra parte, se observa que la clase '*Fearful*' tiende a confundirse con la emoción '*sad*', revelando una dificultad específica en la clasificación posiblemente debido a similitudes en las características emocionales. Además, se identifica un claro desbalance en la clase 2. A pesar de tener menos muestras, la última clase muestra un rendimiento aceptable.

En la segunda iteración de este modelo, **P6V**, se incorpora una validación cruzada al código previo con 5 divisiones (*folds*) y se realiza la evaluación con diversas funciones de activación.

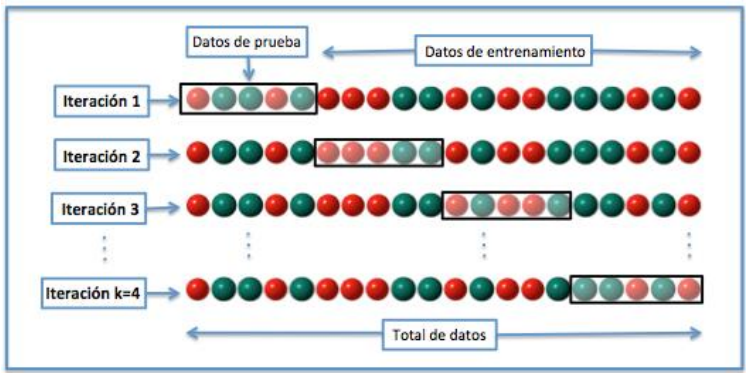


Figura 41. Validación cruzada de K=4 iteraciones [53].

La validación cruzada se utiliza para evaluar el rendimiento de un modelo de aprendizaje automático. En mi caso, se divide el conjunto de datos en 5 subconjuntos o pliegues. Luego, se realiza el entrenamiento y la evaluación del modelo varias veces, utilizando diferentes combinaciones de pliegues como conjuntos de entrenamiento y prueba, después, se promedian los resultados obtenidos en cada iteración para obtener una medida general del rendimiento del modelo. El objetivo es proporcionar una evaluación más robusta asegurando que la red es capaz de generalizar bien las nuevas muestras y no esté memorizando datos del entrenamiento (sobreajuste).

A pesar de las modificaciones, los resultados de esta versión son inferiores a los obtenidos en el modelo anterior.

```
Fold 1: Porcentaje de emociones correctas detectadas en el conjunto de prueba: 17.2543%
Fold 2: Porcentaje de emociones correctas detectadas en el conjunto de prueba: 16.8713%
Fold 3: Porcentaje de emociones correctas detectadas en el conjunto de prueba: 17.8857%
Fold 4: Porcentaje de emociones correctas detectadas en el conjunto de prueba: 16.2393%
Fold 5: Porcentaje de emociones correctas detectadas en el conjunto de prueba: 16.5064%
```

```
Porcentaje de emociones correctas detectadas en el conjunto de entrenamiento: 19.3458%
Porcentaje de emociones correctas detectadas en el conjunto de validación: 17.9082%
Porcentaje de emociones correctas detectadas en el conjunto extra de validación: 19.2%
```

Figura 42. Resultados experimento P6v. (a) Precisión por pliegue. (b) Precisión por conjunto de datos.

En paralelo, se desarrolla un nuevo modelo en **P6z** manteniendo el mismo código de P6 pero esta vez añadiendo la función de activación '*softmax*' en todas las capas, observando mejoras en los resultados:

```
Matriz de Confusión en el conjunto de entrenamiento:
    62    15   250   179   136    35   147    37
     0    32     0    26     6     9     4    58
    13    24   259   129   363    48    84    55
    37    25    95   595    58    35   120    54
    17    32   188    58   612    26    34    44
    19    33   122   423   239    53   101    49
    29    19   162   403   121    61   157    51
    19    29     2    12    11     7     4    76

Precisión en el conjunto de entrenamiento: 33.5135%
>>
Matriz de Confusión en el conjunto extra de validación:
     4     1    23    16     9     6     8     7
     0     1     0     2     0     2     0     3
     2     2    28     8    29     3     8     4
     1     0     6    55    10     6     9     2
     1     2    20     3    44     2     1     6
     1     2     8    26    19     3     6     3
     0     0    17    31    15    15     9     5
     0     1     0     2     0     0     0     3

Precisión en el conjunto de validación extra: 44.4444%
>>
Matriz de Confusión en el conjunto de validación:
    15     8    53    42    27    11    30     5
     0     5     0     3     4     3     0    11
     5     9    61    27    80    10    22    12
     4     5    28   107     9    16    20    11
     2     6    49    13   121     4    12    12
     5     5    27    89    48     7    14    11
     8     6    31    96    32    16    19     7
     7    10     0     5     3     2     1    18

Precisión en el conjunto de validación extra: 32.6087%
>>
```

Figura 43.Resultados experimento P6z.

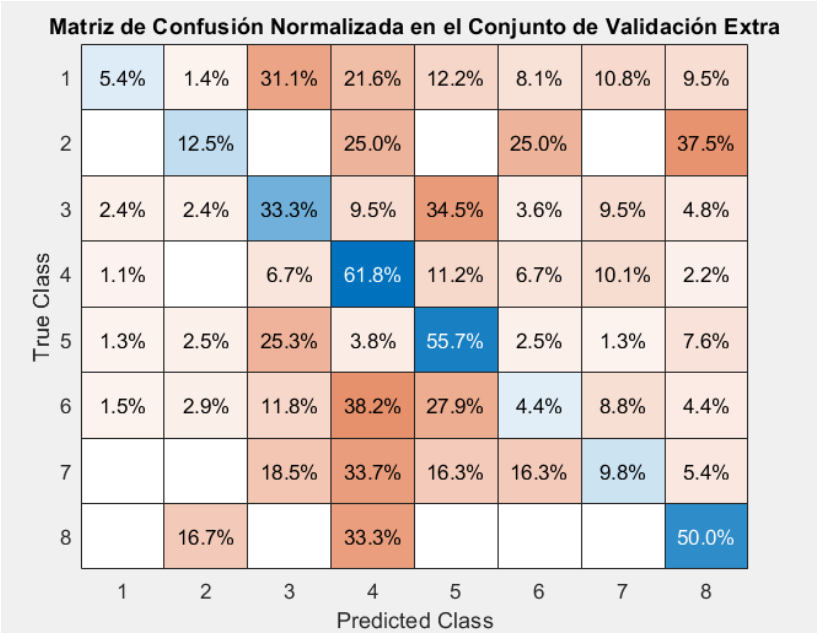


Figura 44. Matriz de confusión normalizada experimento P6z.

Del modelo **P7** en adelante, se juega con nuevas combinaciones de características, se utiliza como referencia el código del entrenamiento P6 con L2 y se añade la entropía y elimina la f0 para probar nuevas opciones.

En **P8** y **P9** se cambia la función de activación de las capas ocultas a RelU (‘*poslin*’) manteniendo ‘*softmax*’ en la capa de salida.

```
%  función de activación de las capas ocultas ReLU
for i=1:numel(hiddenLayerSizes)
    net8.layers{i}.transferFcn = 'poslin'; % ReLU
end

% función de activación de la capa de salida Softmax
net8.layers{end}.transferFcn = 'softmax';
```

Figura 45. funciones de activación experimentos P8 y P9.

Se introduce la técnica de regularización de abandono (*Dropout*) con una tasa del 20%, en **P8** la implementación del *Dropout* se realiza mediante la configuración de la propiedad *Dropout* en net8.performParam para especificar la tasa de abandono.

La función de pérdida utilizada es la entropía cruzad‘ ('*cross-entropy*'), adecuada para problemas de clasificación como al que nos enfrentamos.

```
% regularización dropout
dropoutRate = 0.2; % Tasa de abandono
net8.performFcn = 'crossentropy'; % Función de pérdida para la regularización de abandono
net8.performParam.dropout = dropoutRate;
```

Figura 46. Código regularización Dropout.

Igual que las dos técnicas de regularización explicadas en experimentos anteriores, Dropout sirve para prevenir el sobreajuste. La idea es similar, consiste en ‘apagar’ aleatoriamente un porcentaje de neuronas durante el entrenamiento. Esto significa que, en cada iteración, se ignoran algunas conexiones entre las neuronas para que la red no pueda depender de características específicas en los datos.

En **P9** el Dropout se implementa manualmente en un bucle de entrenamiento, igualmente, desactivando la misma fracción aleatoria de neuronas en cada capa oculta en cada capa:

```
% Dropout manual (similar a p8)
dropoutRate = 0.2; % Tasa de abandono del 20%
for epoch = 1:net9.trainParam.epochs
    % Aplicamos dropout en las capas ocultas
    for i = 1:numel(net9.layers) - 1
        % Para desactivar una fracción aleatoria de neuronas
        if i > 1 % (No aplicamos dropout a la capa de entrada)
            activeIndices = randperm(net9.layers{i}.size, round((1 - dropoutRate) * net9.layers{i}.size));
            net9.layers{i}.userdata.dropoutIndices = false(1, net9.layers{i}.size);
            net9.layers{i}.userdata.dropoutIndices(activeIndices) = true;
        end
    end

    [net9, tr] = train(net9, trainingFeatures', trainingLabelsFeatures');
end
```

.

Figura 47. Implementación Dropout manual en el experimento P9.

De esta forma, se establece la misma tasa de abandono del 20% que representa la fracción de neuronas que se desactivarán durante el entrenamiento.

Se recorre cada capa oculta de la red (sin incluir la capa de entrada para no perder la información importante que contienen las neuronas de esta capa) y desactivan las neuronas de manera aleatoria con la función *randperm*,

Se crea un conjunto de índices 'activeIndices' que representa las neuronas que permanecerán activas y se almacenan los índices de las neuronas desactivadas en el campo *userdata.dropoutindices* de cada capa. Esto permite realizar un seguimiento de qué neuronas están desactivadas en cada época.

Los resultados obtenidos después de aplicar estas técnicas no mejoraron los obtenidos en el entrenamiento **P6z** descrito y graficado anteriormente [Figura 26] y [Figura 27] que alcanzaron una precisión del 33,5% sobre el conjunto de entrenamiento y una precisión del 44,4% sobre el conjunto de validación extra, superando con creces el objetivo inicial puesto en 1/8.

6. CONCLUSIONES

6.2. Conclusiones finales

El desafío en el reconocimiento de emociones a partir de señales de voz radica en la complejidad de abordar la variabilidad de las condiciones de las muestras. Factores como el hablante, el género, el idioma, la duración y la calidad del audio pueden afectar la forma en que se expresan las emociones en la voz. Aunque nosotros, como humanos, somos expertos en esto, trasladar esta habilidad a las máquinas se presenta como un reto ya que se intenta buscar un modelo independiente de estos detalles.

El proyecto realizado confirma la complejidad de la tarea. Se observó que aumentar la cantidad de datos puede beneficiar al modelo, pero también introduce una diversidad que podría afectar su sensibilidad al tener que adaptarse a distintos contextos. La carencia de conjuntos de datos extensos y gratuitos en el ámbito del reconocimiento de emociones de voz es un problema evidente. Aunque la disponibilidad de más datos podría ser ventajosa, no garantiza una mejora sustancial en el rendimiento del modelo.

Cabe destacar que la mayoría de las bases de datos existentes presentan limitaciones para evaluar el rendimiento, ya que incluso para los humanos, determinar las emociones en algunas grabaciones seleccionadas puede ser un proceso muy complicado.

En este punto, se extraen diversas conclusiones. La evaluación de la precisión del modelo, su capacidad de generalización a datos no vistos, y la comparación de rendimiento entre diferentes bases de datos ofrecen percepciones sobre la robustez y la adaptabilidad del modelo a distintos contextos. Además, el análisis de sensibilidad a la calidad de las señales de voz, la optimización de hiperparámetros, y la exploración de errores cometidos proporcionan información valiosa para mejorar el modelo. La interpretación de las características que influyen en la clasificación de emociones y la comparación con otros enfoques contribuyen a una comprensión más completa del rendimiento del sistema de reconocimiento de emociones basado en señales de voz.

No obstante, a pesar de haber mejorado los resultados esperados desarrollando un modelo que alcanzó una precisión del 44,40% en el conjunto de datos de validación extra, superando significativamente el umbral mínimo esperado del 12,5% de precisión, establecido como referencia para un clasificador aleatorio de ocho emociones, se considera que el modelo actual podría ser más robusto. Por esto, en el siguiente apartado, se proponen nuevas mejoras, líneas de investigación y enfoques alternativos con el objetivo de elevar la eficacia del sistema de reconocimiento de emociones.

6.3. Competencias empleadas

La formación adquirida durante la carrera en asignaturas de programación como Informática I, Informática II y Protocolos, ha sido esencial para la comprensión y desarrollo de redes neuronales. Estas asignaturas proporcionaron una base sólida y conocimientos fundamentales que han resultado muy útiles a la hora de abordar conceptos más complejos en el ámbito de la inteligencia artificial.

Además, asignaturas como Estándares o Equipos, me han permitido profundizar en el análisis de señales de audio, especialmente a través del uso de Matlab. He podido también recurrir a mis apuntes de asignaturas como Sistemas Lineales y Fundamentos, toda esta formación me ha preparado para enfrentarme a problemas más complicados, aplicando fórmulas estadísticas avanzadas y realizando operaciones con señales, incluyendo técnicas de Fourier, entre otras.

6.4. Competencias adquiridas

A lo largo de mi Trabajo de Fin de Grado he adquirido diversas habilidades que han enriquecido mi viaje académico y personal. Aquí destaco las competencias que han marcado mi desarrollo:

En primer lugar, utilizando MATLAB para desarrollar un modelo de aprendizaje automático he explorado a fondo este entorno y además de

mejorar mi destreza en programación, también me ha permitido descubrir nuevas posibilidades en el mundo de la inteligencia artificial y sumergirme en el reconocimiento vocal de emociones, que no solo ha ampliado mi conocimiento técnico, sino que me ha proporcionado otra visión de los aspectos psicológicos y neurocientíficos vinculados a la expresión emocional a través de la voz.

Por otra parte, he perfeccionado la habilidad de extraer información relevante de señales de audio, centrándome en las características más importantes en la detención de emociones y he llevado a la práctica conocimientos adquiridos durante los 4 años de carrera sobre el procesamiento de señales e implementación, entrenamiento y optimización de redes neuronales.

6.5. Trabajos futuros

A continuación, se plantean futuras mejoras y líneas de investigación que considero que podrían enriquecer mi proyecto.

- **Estudio profundo de características físicas de la voz.** Se plantea realizar un estudio previo sobre las características físicas y acústicas de la voz en busca de comprender como varían en los diferentes estados emocionales y como se pueden utilizar para mejorar la detención de emociones.
- **Biblioteca SHAP (SHapley Additive exPlanations).** Se propone realizar un análisis más detallado de las características utilizadas para entrenar la red neuronal. La utilización de SHAP en Python, por ejemplo, permitirá calcular la relevancia de las características, visualizando la contribución de cada una de estas y cómo afecta a las predicciones del modelo.
- **Modelo BERT (Bidirectional Encoder Representations from Transformers).** Con la incorporación de BERT en el análisis de sentimientos se buscaría aprovechar la habilidad de este modelo para comprender el contexto y las relaciones entre palabras, podría mejorar la precisión del análisis al capturar matices emocionales más complejos.

- **Técnicas de Análisis de Sentimientos.** Se sugieren enfoques más avanzados como la integración de Transformers como evolución de las redes LSTM. Estos modelos ofrecen una paralelización para procesar información simultáneamente que podría potenciar los entrenamientos con conjuntos de datos más extensos, mejorando así la captura de matices en el habla.
- **Optimización de Procesamiento.** Aumentar la mejora del rendimiento del modelo mediante la implementación de hardware especializado, como GPUs o aprovechando plataformas en la nube para un procesamiento más eficiente y rápido.
- **Exploración de clasificadores.** Aunque las arquitecturas existentes sean robustas y se haya comprobado que son capaces de realizar esta tarea, la elección de un clasificador óptimo sigue siendo un desafío dada la variabilidad de los datos. Se propone por esto continuar explorando y evaluando con el fin de perfeccionar la elección de clasificadores que optimicen el rendimiento en diversas condiciones y contextos.
- **Diversificación de bases de datos**. Se sugiere investigar y experimentar con bases de datos que incluyan señales de voz con contenido emocional espontáneo. El uso de señales de audio 'naturales' como conversaciones diarias o grabaciones auténticas podría mejorar la capacidad del modelo para reconocer y adaptarse a una variedad más amplia de expresiones.
- **Reducción del número de emociones a detectar.** Se plantea la opción de probar con menos clases en lugar de con las ocho utilizadas en el proyecto, con el objetivo de evaluar si esta simplificación contribuye a un mejor rendimiento del modelo.

7. PRESUPUESTO

La licencia estándar de MATLAB tiene un costo de 900 euros al año [52]. En términos de medios físicos, se requiere un ordenador capaz de ejecutar este programa, gestionar bases de datos y realizar numerosos entrenamientos. En mi caso, Utilicé 14 GB de memoria para desarrollar el proyecto. Un ordenador de gama media tiene un costo aproximado de 700 euros.

Además, es necesario considerar los costos asociados con las horas dedicadas por los investigadores. El proyecto tiene asignados 12 créditos ETCS, lo que, según el plan de estudios del grado, equivale a 25 horas por crédito [54], sumando un total de 300 horas. Considerando que un ingeniero Junior cobra 12e/h [55] y añadiendo el trabajo del tutor, quien podría cobrar 20e/h, el costo aproximado podría calcularse de la siguiente manera:

Ctemporal = Número de créditos ∗ Horas por crédito ∗ (Tarifa ingeniero junior + Tarifa tutor)

Ctemporal = 12 x 25 x (12 + 20) = 9.600 euros

Si a esto le sumamos la licencia y el ordenador, el coste total del proyecto es el siguiente:

Ctotal = Ctemporal + C MATLAB + C PC

Ctotal = 9.600 + 900 + 700 = 11.200 euros

Tipo de Coste	Coste Total (euros)
Mano de obra	9.600
Licencias	900
Materiales	700
TOTAL	11.200

8. REFERENCIAS BIBLIOGRÁFICAS

[1] el-perruco. (2019, julio 29). *Emociones básicas: Qué y cuáles son las emociones primarias y secundarias*. el perruco; "el perruco". https://www.elperruco.com/emociones-basicas-que-cuales-son-emociones-primarias-secundarias/

[2] Wikipedia contributors. (s/f). Robert Plutchik. Wikipedia, The Free Encyclopedia. https://es.wikipedia.org/w/index.php?title=Robert_Plutchik&oldid=153290561

[3] Vázquez, K. (2016, April 1). *Tu ordenador sabrá si estás triste*. Muy Interesante. https://www.muyinteresante.com/curiosidades/13898.html

[4] Vázquez, K. (2016b, May 17). *Llegan las máquinas que detectan emociones*. Muy Interesante. https://www.muyinteresante.com/tecnologia/14782.html

[5] Huang, X., Acero, A., & Hon, H. (2001). Spoken language processing: A guide to theory, algorithm, and system development. Prentice-Hall.

[6] Saa-Vernaza, Á. J., & Mosquera, E. F. (2018). *Razonamiento Covariacional al estudiar la función por partes mediado por GeoGebra*. https://humanidades.com/aparato-fonador/

[7] *Cuidando su voz. (n.d.). NIDCD. Retrieved February 29, 2024, from https://www.nidcd.nih.gov/es/espanol/cuidando-su-voz*

[8] Calvo, D. (2020, March 1). *La voz y las emociones*. Locutora Online; Diana Calvo. https://vozyemociones.com/la-voz-y-las-emociones/

[9] Zhang, B., Essl, G., & Provost, E. M. (2015, September). Recognizing emotion from singing and speaking using shared models. In 2015 international conferen- ce on affective computing and intelligent interaction (acii) (pp. 139-145). IEEE.

[10] Petrushin, V. *"Emotion in Speech: Recognition and Application in Call Centers," Neural Network Engineering Artefact, 2000, 710, 22.*

[11] Nakatsu, R., Nicholson, J., & Tosa, N. *(1999, October). Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In Proceedings of the seventh ACM international conference on Multimedia (Part 1) (pp. 343-351).*

[12] Ashish B. Ingale, D. S. Chaudhari. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012. Speech Emotion Recognition

[13] Sidorov, M., Ultes, S., & Schmitt, A. *(2014, May). Emotions are a personal thing: Towards speaker-adaptive emotion recognition. In 2014 IEEE international*

conference on acoustics, speech, and signal processing (ICASSP) (pp. 4803-4807). IEEE.

[14] Jurado, F. J. A., Vadillo, E. D., & Massana., E. F. I. (n.d.). *PATOLOGÍA DE LA VOZ HABLADA Y DE LA VOZ CANTADA*. Seorl.net. Retrieved February 20, 2024, from https://seorl.net/PDF/Laringe%20arbor%20traqueo-bronquial/117%20-%20PATOLOG%C3%8DA%20DE%20LA%20VOZ%20HABLADA%20Y%20DE%20LA%20VOZ%20CANTADA.pdf

[15] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. *(2019). Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7, 117327-117345.*

[16] Boole, G. (1854). An Investigation of the Laws of Thought. Walton and Maberly.

[17] Turing, A. M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, 42(2), 230-265.

[18] Wikipedia contributors. (n.d.-b). *Deep Blue (computadora)*. Wikipedia, The Free Encyclopedia. https://es.wikipedia.org/w/index.php?title=Deep_Blue_(computadora)&oldid=157477303

[19] Josep. (2023, July 2). *Diferencia entre aprendizaje supervisado y no supervisado: ¿Cuál es la mejor opción?* Conectando ideas; Josep. https://conectandoideas.net/diferencia-aprendizaje-supervisado-y-no-supervisado/

[20] *Aprendizaje automático supervisado - Predicciones en ML*. (n.d.). CodeToDevs | Python, Django, CSS, JavaScript. Retrieved September 12, 2023, from https://www.codetodevs.com/aprendizaje-automatico-supervisado/

[21] *Deep Learning o Aprendizaje profundo: ¿qué es?* (2022, April 19). Formación en ciencia de datos | Datascientest.com; DataScientest. https://datascientest.com/es/deep-learning-definicion

[22] Jasim, Y., Saeed, M. G., & Raewf, M. (2023). Analyzing social media sentiment: Twitter as a case study. *ADCAIJ ADVANCES IN DISTRIBUTED COMPUTING AND ARTIFICIAL INTELLIGENCE JOURNAL*. https://doi.org/10.14201/adcaij.28394

[23] *¿Qué son las redes neuronales?* (n.d.). Ibm.com. Retrieved February 20, 2024, from https://www.ibm.com/mx-es/topics/neural-networks

[24] de Grado, T. F. (n.d.). *CLASIFICACIÓN DE HUEVOS POR TAMAÑO USANDO REDES NEURONALES*. Umh.Es. Retrieved September 12, 2023, from http://dspace.umh.es/bitstream/11000/7685/1/TFG-Pomares%20Palomares%2C%20Teresa.pdf

[25] (N.d.-c). Researchgate.net. Retrieved February 20, 2024, from https://www.researchgate.net/figure/Common-neural-network-activation-functions_fig7_305881131

[26] Shopenova, A. (2021, March 7). *Predict customer churn with neural network*. Towards Data Science. https://towardsdatascience.com/predict-customer-churn-with-neural-network-1ef8f1a1c6ab

[27] Camejo Corona, J., Gonzalez, H., & Morell, C. (2019). Los principales algoritmos para regresión con salidas múltiples. Una revisión para Big Data. *Revista Cubana de Ciencias Informáticas*, *13*(4), 118–150. http://scielo.sld.cu/scielo.php?pid=S2227-18992019000400118&script=sci_arttext

[28] (N.d.-d). Amazon.com. Retrieved September 12, 2023, from https://aws.amazon.com/es/what-is/neural-network/

[29] de Grado, T. de F. (n.d.). *Emociones en Señales de Voz: Reconocimiento con Redes Neuronales Profundas*. Upc.edu. https://upcommons.upc.edu/bitstream/handle/2117/363290/Memoria_TFG..pdf?sequence=2

[30] *¿Qué son las redes neuronales recurrentes?* (n.d.). Ibm.com. Retrieved https://www.ibm.com/es-es/topics/recurrent-neural-networks

[31] *Redes Neuronales Recurrentes*. (n.d.). Torres.ai. Retrieved September 12, 2023, , from https://torres.ai/redes-neuronales-recurrentes/

[32] Chauhan, A. (2021, May 15). *Why LSTM more useful than RNN in Deep Learning?* Towards AI. https://pub.towardsai.net/deep-learning-88e218b74a14

[33] Lapedriza, À. (2019, April 1). *Computación afectiva. ¿Un robot puede tener empatía?* Tecnología++. https://blogs.uoc.edu/informatica/computacion-afectiva/

[34] Banafa, A. (2016, June 6). *¿Qué es la computación afectiva?* OpenMind. https://www.bbvaopenmind.com/tecnologia/mundo-digital/que-es-la-computacion-afectiva/

[35] Vives, V. (2019, January 8). *5 aplicaciones para trabajar la inteligencia emocional*. Blog Vicens Vives. https://blog.vicensvives.com/5-apps-para-trabajar-la-inteligencia-emocional/

[36] Coolhuntermx, R. (2021, August 24). *4 apps para la salud mental y emocional*. Coolhuntermx. https://coolhuntermx.com/apps-de-inteligencia-artificial-para-la-salud-mental-y-emocional/

[37] Javier De Lope. et al. A Hybrid Time-Distributed Deep Neural Architecture for Speech Emotion Recognition. *World Scientific 2022*. https://doi.org/10.1142/S0129065722500241

[38] AAZG. (2023, June 18). *🎾 Dominando la Matriz de Confusión: La guía completa para entender el rendimiento de nuestros modelos de Machine Learning 🚀*. Medium. https://medium.com/@aazg24/dominando-la-matriz-de-confusi%C3%B3n-la-gu%C3%ADa-completa-para-entender-el-rendimiento-de-nuestros-39cfcc53ddd5

[39] de los Santos, P. R. (2021, December 13). Cómo interpretar la matriz de confusión: ejemplo práctico. *Telefónica Tech*. https://telefonicatech.com/blog/como-interpretar-la-matriz-de-confusion-ejemplo-practico

[40] Wikipedia contributors. (n.d.-b). *Curva ROC*. Wikipedia, The Free Encyclopedia. https://es.wikipedia.org/w/index.php?title=Curva_ROC&oldid=155304929

[41] *Vista de Análisis de sentimientos para Twitter con Vader y TextBlob. (n.d.)*. *Edu.ec. Retrieved March 11, 2023, from https://revista.uisrael.edu.ec/index.php/ro/article/view/494/437*

[42] Livingstone, S. R. (2019). *RAVDESS Emotional speech audio* [Data set].

[43] Sunkaraneni, T. (2019). *SAVEE Database* [Data set].

[44] Lok, E. J. (2019). CREMA-D [Data set].

[45] Wikipedia contributors. (n.d.-b). *Conjuntos de datos de entrenamiento, validación y prueba*. Wikipedia, The Free Encyclopedia. https://es.wikipedia.org/w/index.php?title=Conjuntos_de_datos_de_entrenamiento,_validaci%C3%B3n_y_prueba&oldid=157110805

[46] *Los sets de entrenamiento, validación y prueba*. (n.d.). Codificando Bits. Retrieved February 20, 2024, from https://www.codificandobits.com/blog/sets-entrenamiento-validacion-y-prueba/

[47] Lathrop, R, (2019). Introduction to machine learning improve performance by observation from https://slidetodoc.com/

[48] *Deep Learning: clasificando imágenes con redes neuronales*. (2020, May 25). LIS Data Solutions. https://www.lisdatasolutions.com/es/blog/deep-learning-clasificando-imagenes-con-redes-neuronales/

[49] *stft*. (n.d.). Mathworks.com. Retrieved January 21, 2024, from https://la.mathworks.com/help/signal/ref/stft.html

[50] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[51] R. W. Picard, "Affective Computing, *"MIT Media Laboratory; Perceptual Computing 20 Ames St., Cambridge.*

[52] (N.d.-d). Quizlet.com. Retrieved February 23, 2024, from https://quizlet.com/562645140/dermheent-wk-1-yr2-stridor-phonation-and-mucous-membrane-disorders-flash-cards/

[53] Wikipedia contributors. (n.d.-f). *Validación cruzada*. Wikipedia, The Free Encyclopedia. https://es.wikipedia.org/w/index.php?title=Validaci%C3%B3n_cruzada&oldid=155368779

[53] *Pricing and licensing*. (n.d.). Mathworks.com. Retrieved February 27, 2024, from,https://es.mathworks.com/pricing licensing.html?prodcode=ML&intendeduse=comm

[54] (N.d.-e). Gob.Es. Retrieved February 27, 2024, from https://www.educacionyfp.gob.es/italia/dam/jcr:b53864d2-65a3-4526-abf4-61ef02f5be34/el-sistema-universitario-espa-ol2.pdf

[55] Salario para Ingeniero Junior en España - Salario Medio. (n.d.). Talent.com. Retrieved February 27, 2024, from https://es.talent.com/salary?job=ingeniero+junior

ANEXO

El código fuente desarrollado para este proyecto se encuentra disponible en el siguiente repositorio público en GitHub.

https://github.com/mariamrtz00/Audio-Emotion-Recognition-System.git

Este repositorio contiene la implementación completa del sistema, incluyendo todos los scripts, archivos de configuración y recursos utilizados durante el desarrollo. Este enfoque permite a los interesados examinar el código, comprender la implementación y colaborar en el desarrollo continuo del proyecto.

A continuación, se presentan algunos aspectos destacados del contenido del repositorio:

Se incluye un detallado README con una breve descripción del proyecto y enlaces para descargar las bases de datos utilizadas. También se adjunta la memoria del Trabajo de Fin de Grado en formato PDF, junto con una tabla de Excel que presenta de manera exhaustiva los resultados de los entrenamientos más significativos, incluyendo observaciones más detalladas. Además, incluye los archivos .m mencionados en la memoria, junto con sus correspondientes redes neuronales ya entrenadas.