Análisis de voz con inteligencia artificial para identificar las emociones de los estudiantes en una universidad pública

21 Julio 2025

Universidad Internacional de Valencia

Titulación:

Maestría en Inteligencia Artificial

Curso académico

2024 – 2025

Alumno:

Flores Masías, Edward José

D.N.I: 09536323

Directora de TFM: Jimena Llopis Rivas

Convocatoria:

Segunda

De:
Planeta Formación y Universidades

Índice

Índice de ilustraciones

Índice de tablas

Universidad Internacional de Valencia

Resumen

El presente trabajo ha sido desarrollado en la ciudad de Lima-Perú, con el propósito de identificar las emociones de los estudiantes ingresantes en una universidad pública, tiene como propósito digitalizar la señal acústica, para que pueda ser tratada como datos y a partir de ahí, que se puede generar un modelo en inteligencia artificial para identificar principalmente las emociones de alegría, tristeza, enojo, neutral, sorpresa y miedo. La primera parte del del trabajo consiste en realizar un estado del arte sobre las fuentes relevantes que permitieron desarrollar el propósito del estudio, en esta parte de identifica el comportamiento de la voz humana y los elementos para medir su comportamiento, como la frecuencia, longitud de onda, etc. Posterior a ello, se presenta una revisión conceptual de las diversas formas y técnicas del comportamiento de la voz humana a partir de la digitalización, gran parte de este aporte fue el desarrollo de las escalas o coeficientes cepstrales de Mel (MFCC), que dividen la señal de audio en pequeños segmentos para obtener su espectro de potencia. El espectro de frecuencia se pasa a través de los filtros de Mel, que se encargan de distribuirse de acuerdo con su escala y que se asemeja a la forma en que el oído humano percibe las frecuencias. Posterior a ello, se hace una revisión de los diferentes tipos de red neuronal para trabajar e identificar cuál sería la más ideal. En la etapa de la metodología, se describe cuál será el proceso a seguir para el estudio, identificando las capacidades y necesidades de los filtros a aplicar, así como la limpieza y consideraciones de los datos recogidos. Posteriormente, en la etapa de desarrollo se elabora un programa en Python que permite generar el modelo de reconocimiento de emociones, logrando así cumplir con el objetivo propuesto.

Palabras clave: MFCC, inteligencia artificial, CNN, análisis de sentimientos.

Abstract

This work was developed in Lima, Peru, with the purpose of identifying the emotions of incoming students at a public university. Its purpose is to digitize the acoustic signal so that it can be processed as data. From there, an artificial intelligence model can be generated to identify primarily the emotions of joy, sadness, anger, neutrality, surprise, and fear. The first part of the work consists of conducting a state-of-the-art analysis of the relevant sources that allowed for the development of the study's purpose. In this part, the behavior of the human voice and the elements to measure its behavior, such as frequency, wavelength, etc., are identified. Following this, a conceptual review of the various forms and techniques of human voice behavior based on digitization is presented. A large part of this contribution was the development of Mel cepstral scales or coefficients (MFCC), which divide the audio signal into small segments to obtain its power spectrum. The frequency spectrum is passed through Mel filters, which distribute the frequency spectrum according to its scale, resembling the way the human ear perceives frequencies. Following this, a review of the different types of neural networks is conducted to identify the most ideal one. In the methodology phase, the study process is described, identifying the capabilities and requirements of the filters to be applied, as well as the cleaning and considerations of the collected data. Subsequently, in the development phase, a Python program is created to generate the emotion recognition model, thus achieving the proposed objective.

Keywords: MFCC, artificial intelligence, CNN, sentiment analysis.

1. Introducción

La identificación y evaluación de las emociones en estudiantes universitarios constituyen un área de creciente interés en los campos de la educación y la psicología, particularmente en un contexto donde las emociones desempeñan un papel fundamental en el proceso de aprendizaje. La inteligencia artificial se ha consolidado como un instrumento eficaz para facilitar la evaluación y comprensión de dichas emociones mediante el análisis de la voz, un elemento inherente de la comunicación humana. A través de algoritmos sofisticados y técnicas de aprendizaje automático, se pueden identificar fluctuaciones en el tono, el ritmo y el timbre vocal que pueden desvelar estados emocionales latentes, proporcionando así una comprensión más profunda del bienestar estudiantil.

La intersección de la inteligencia artificial y la inteligencia emocional ha cobrado cada vez más importancia en los últimos años, especialmente en contextos educativos donde comprender las emociones de los estudiantes puede influir profundamente en los resultados de aprendizaje. Los avances recientes en la tecnología de análisis de voz han abierto nuevas vías para desentrañar las complejidades de la expresión emocional en los estudiantes (Yogesh K. Dwivedi et al., 2022). Además, a medida que los sistemas educativos priorizan cada vez más el aprendizaje socioemocional, el papel del análisis de voz se vuelve crucial para promover un enfoque holístico de la educación que aborde no solo el desarrollo cognitivo sino también el emocional. Un tema clave que emerge de la literatura actual es la eficacia de las diferentes metodologías de inteligencia artificial para reconocer e interpretar las señales emocionales a partir de datos vocales. Diversos estudios demuestran la eficacia de los algoritmos de aprendizaje automático para distinguir diversos estados emocionales según características de la voz, como el tono, el timbre y el ritmo (Mohammed YA, 2024). Si bien los estudios han destacado las implicaciones positivas del uso de la IA para mejorar la conciencia emocional del alumnado, la evidencia empírica que respalda los resultados a largo plazo sigue siendo escasa, lo que exige estudios longitudinales que evalúen el impacto del análisis de voz en el rendimiento y el crecimiento emocional del alumnado a lo largo del tiempo (Gilleran et al., 2019).

Además, el interés por investigar el efecto de las emociones en el rendimiento académico ha propiciado la incorporación de técnicas de análisis de voz en investigaciones de mayor alcance en el contexto educativo. Conforme se optimiza la precisión de los modelos de inteligencia artificial para la identificación emocional, resulta relevante explorar cómo estas tecnologías pueden funcionar como instrumentos auxiliares para psicólogos, educadores e investigadores. Esta investigación se justifica no solo por la necesidad de tratar el bienestar emocional de los alumnos, sino también por su potencial para modificar la dinámica de enseñanza-aprendizaje, incorporando la dimensión emocional como un componente esencial en la formación holística del estudiante universitario.

2. Objetivos

El presente Trabajo de Fin de Master propone el desarrollo de un modelo inteligente capaz de identificar y analizar las emociones de los estudiantes mediante el procesamiento de señales de voz, lo cual representa un avance significativo frente a los métodos tradicionales. Esta solución busca brindar a los docentes herramientas tecnológicas que les permitan comprender mejor el estado emocional de sus alumnos y adaptar sus metodologías de enseñanza de manera más efectiva.

Además, el desarrollo de una aplicación de recolección y análisis emocional contribuirá al fortalecimiento de entornos de aprendizaje más empáticos, inclusivos y centrados en el estudiante. La implementación de este tipo de sistemas no solo mejora la interacción docente-estudiante, sino que también favorece la toma de decisiones pedagógicas basadas en evidencia.

Por lo tanto, este proyecto responde a una necesidad real en el ámbito educativo y representa una contribución significativa tanto a nivel académico como tecnológico, con potencial de aplicación en distintas instituciones de educación superior, para lo cual, se plantean los siguientes objetivos.

OBJETIVO GENERAL

Desarrollar un sistema basado en inteligencia artificial que analice la voz de los estudiantes para identificar sus emociones en el contexto de una universidad pública.

OBJETIVOS ESPECÍFICOS

- Recolectar un conjunto de datos de voz etiquetado por emociones representativas del contexto académico universitario.
- Implementar algoritmos de procesamiento de voz en Python para la extracción de características.
- Entrenar un modelo de aprendizaje automático y aprendizaje profundo para la clasificación de emociones a partir de los patrones vocales.
- Evaluar el rendimiento de los modelos desarrollados mediante métricas como precisión, recall, F1-score y exactitud, usando validación cruzada.

3. Estado del Arte y Marco teórico

El análisis de voz se define como una técnica que examina características acústicas del habla para extraer información relevante sobre las emociones y estados psicológicos del hablante. Este proceso implica la utilización de tecnología avanzada que permite cuantificar variables como la frecuencia fundamental y la variabilidad de la voz, las cuales son indicadores significativos del nivel de tensión emocional experimentada por un individuo. Bajo este enfoque, la tecnología de agentes empáticos, tal como se describe en las innovaciones recientes, se basa en medidas indirectas de la respuesta fisiológica asociada al estrés, evidenciando la importancia del análisis de la voz en la comprensión de las emociones humanas (Broek et al., 2005). Además, la integración de enfoques de medicina tradicional, como el Qigong, con el análisis de voz mediante inteligencia artificial, resalta un avance hacia la creación de herramientas que pueden prever patrones emocionales y guiar intervenciones médicas personalizadas (Dethlefs et al., 2019).

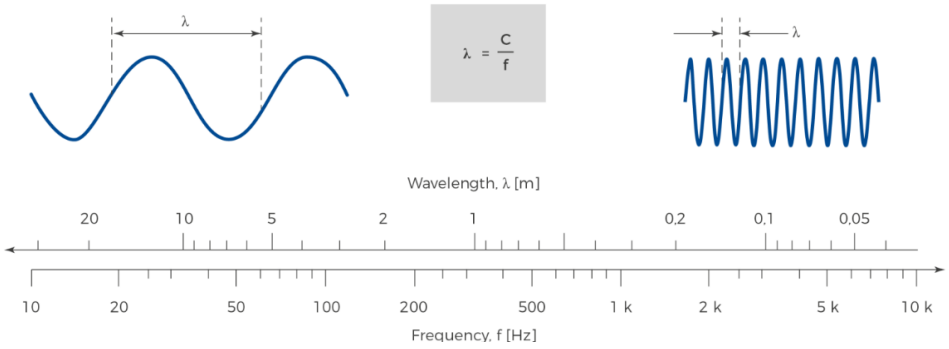


Ilustración 1. Longitud y frecuencia de una onda (Svantek, 2025).

El análisis de voz mediante inteligencia artificial se ha convertido en un campo crucial para la identificación de emociones, y su evolución marca un avance significativo en las interacciones humano-máquina. Esta tecnología permite descomponer la voz en sus componentes fundamentales, como el tono, la entonación y los patrones de habla, para captar matices emocionales que antes eran difíciles de detectar. En el contexto actual, la investigación sugiere que los agentes artificiales no solo muestran expresividad emocional, sino que también pueden evocar respuestas emocionales en los seres humanos, un fenómeno que está profundamente vinculado con la percepción emocional (Cross et al., 2018). Además, eventos como el de recuperación de información han subrayado la importancia de integrar la inteligencia artificial en soluciones de análisis vocal, promoviendo diálogos espaciales entre expertos en diversas disciplinas (Azzopardi et al., 2017). Así, el análisis de voz se posiciona como un puente vital entre la tecnología y la experiencia emocional humana.

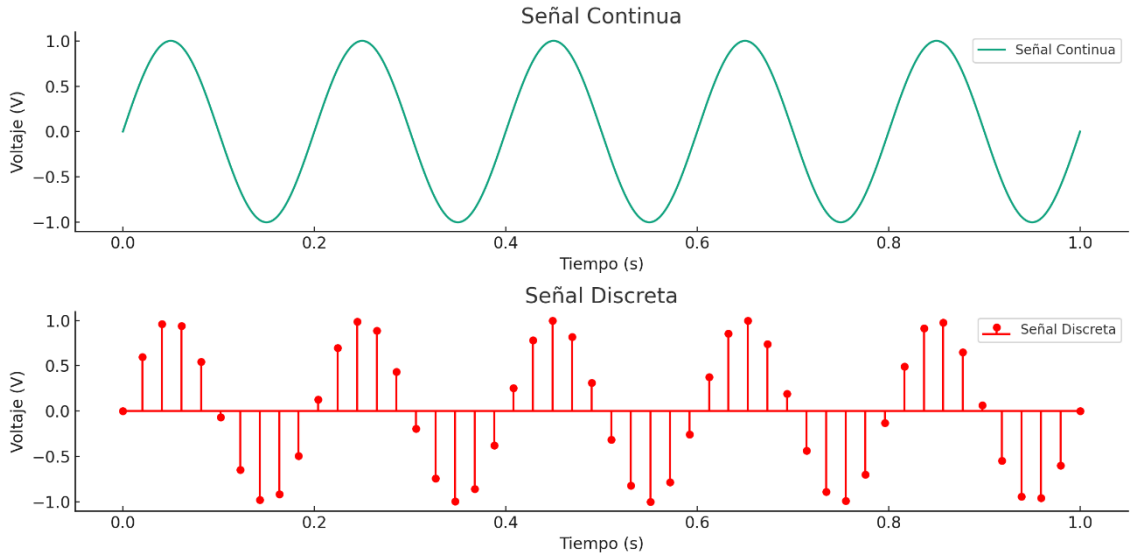


Ilustración 2. Diferencia entre la señal continua y discreta (García, 2023).

La exploración del análisis de voz mediante inteligencia artificial (IA) para identificar las emociones de los estudiantes revela una comprensión multifacética de la dinámica emocional en entornos educativos. Investigadores han establecido que los sistemas basados en IA, que aprovechan algoritmos avanzados de aprendizaje automático, pueden discernir eficazmente diversos estados emocionales basándose en características vocales, como el tono, el timbre y el ritmo. Esta capacidad de reconocimiento emocional matizado no es meramente académica; tiene implicaciones prácticas para mejorar la participación estudiantil, el rendimiento académico y fomentar entornos educativos más receptivos (Mohammed YA, 2024). A medida que los marcos educativos priorizan cada vez más el aprendizaje socioemocional, la capacidad de la tecnología de análisis de voz para identificar e interpretar señales emocionales ofrece un medio fundamental para que los educadores adapten sus estrategias pedagógicas y satisfagan las necesidades emocionales de los estudiantes (Azzopardi et al., 2017). Un aspecto central de la revisión ha sido el reconocimiento de las consideraciones éticas en torno a las aplicaciones de la IA en la psicología educativa. Los estudios destacan la persistente preocupación por la privacidad y la seguridad de los datos, en particular dada la naturaleza sensible de la información emocional que se procesa (Gilleran et al., 2019). A medida que estas tecnologías avanzan, se hace evidente la apremiante necesidad de marcos éticos sólidos, lo que impulsa un diálogo sobre cómo equilibrar los beneficios de las perspectivas impulsadas por la IA con el imperativo de salvaguardar los datos emocionales de los estudiantes (Sivek et al., 2018) (Grawemeyer et al., 2017). Además, a pesar de los prometedores avances observados en la literatura actual, existen limitaciones significativas con respecto al enfoque demográfico de los estudios existentes, que a menudo priorizan a las poblaciones adultas sobre los estudiantes (Soelistio et al., 2017). Esta observación señala la

necesidad de más investigación que atienda específicamente a la diversidad demográfica de los estudiantes para garantizar que la tecnología diseñada para la evaluación emocional sea adecuadamente matizada y contextualizada.

A medida que los sistemas de IA se han vuelto más sofisticados, la capacidad de estas herramientas para interpretar no solo el contenido verbal, sino también las emociones subyacentes han cobrado relevancia. Este enfoque interdisciplinario, que integra la psicología, la lingüística y la ingeniería de software, potencia aplicaciones que van desde la atención al cliente hasta el diagnóstico médico, lo que convierte la identificación emocional en una habilidad valiosa en contextos sociales y comerciales. La investigación previa ha demostrado que las variaciones en el tono, la velocidad y el patrón del habla son indicadores fundamentales de las emociones humanas, permitiendo que algoritmos aprendan a reconocer estas señales con un alto grado de precisión (Najjar R, 2023). Además, el desarrollo de técnicas de aprendizaje profundo ha mejorado significativamente la eficacia de estos sistemas, generando un incremento en la investigación centrada en la aplicación de redes neuronales para el análisis emocional (Kamalov F et al., 2023). La relevancia del análisis de emociones a través de la voz se ha reflejado en numerosos estudios que destacan su utilidad en entornos donde la comunicación efectiva es crítica. Por ejemplo, en el ámbito de la salud mental, una adecuada identificación de las emociones a través del habla puede facilitar intervenciones más efectivas y personalizadas (Yogesh K Dwivedi et al., 2023). Del mismo modo, en el sector empresarial, comprender las emociones de los clientes puede resultar en una mejora en la experiencia del usuario y, por ende, en la fidelización (Yang Q et al., 2022). A pesar de estos avances, hay un conjunto significativo de desafíos y brechas en la literatura existente que requieren atención. La mayoría de los estudios se centran en un número limitado de idiomas y contextos culturales, lo que plantea preguntas sobre la aplicabilidad de los modelos entrenados en un contexto a otro (Yogesh K Dwivedi et al., 2021). Además, los enfoques actuales a menudo dependen de bases de datos que pueden no ser representativas de la diversidad de expresiones emocionales presente en la población general (Varnosfaderani SM et al., 2024). Otro aspecto que emerge de la revisión de la literatura es la necesidad de considerar factores como la tonalidad, el acento y los elementos no verbales que pueden influir en la interpretación emocional (Bankins S et al., 2023). Si bien algunos estudios han comenzado a abordar estas cuestiones (Budhwar P et al., 2023), aún falta una comprensión más profunda de cómo interactúan estos elementos para crear un análisis emocional más holístico. Por ende, es esencial implementar investigaciones que aborden la variabilidad del lenguaje y otras características socioculturales que afectan la expresión de emociones (Kuwaiti AA et al., 2023). También se ha observado que, aunque hay esfuerzos significativos en el desarrollo de modelos predictivos, la validación de estos modelos en condiciones del mundo real sigue siendo un aspecto crítico que no ha recibido suficiente atención (Yogesh K Dwivedi et al., 2022).

La transición hacia enfoques más sofisticados se evidenció con la introducción de técnicas de procesamiento de lenguaje natural que, según (Yang Q et al., 2022),

ampliaron las capacidades analíticas al incluir no solo el contenido verbal, sino también el contexto emocional del habla. Este cambio marcó un hito en la evolución del análisis de emociones, ya que las emociones complejas, como la ironía o la tristeza sutil, comenzaron a ser más accesibles para la detección automática (Yogesh K Dwivedi et al., 2021), (Varnosfaderani SM et al., 2024). Con el paso del tiempo, se han desarrollado modelos más robustos que integran redes neuronales profundas, lo que ha permitido un reconocimiento emocional más matizado y adaptable a distintas situaciones comunicativas (Bankins S et al., 2023), (Budhwar P et al., 2023). En este sentido, recientes investigaciones (Kuwaiti AA et al., 2023), (Yogesh K Dwivedi et al., 2022) subrayan la convergencia de la tecnología de análisis de voz y la psicología del comportamiento, sugiriendo un marco interdisciplinario cada vez más necesario para abordar la complejidad emocional humana de manera efectiva.

Un aspecto relevante en la literatura es el uso de características acústicas como indicadores de estados emocionales. Diferentes estudios han encontrado que parámetros como la variación tonal y el ritmo del habla son significativos en la clasificación de emociones (Yogesh K Dwivedi et al., 2023), (Yang Q et al., 2022). Además, la contextuación cultural del análisis de emociones a través de la voz ha sido objeto de discusión, sugiriendo que la interpretación de emociones puede variar sustancialmente entre diferentes entornos culturales (Yogesh K Dwivedi et al., 2021), (Varnosfaderani SM et al., 2024).

La integración de técnicas de procesamiento del lenguaje natural con el análisis de voz ha permitido avances importantes en la identificación de emociones complejas. Esto ha sido corroborado por diversas investigaciones que demuestran cómo los modelos de aprendizaje profundo pueden mejorar la precisión del reconocimiento emocional, apoyando la premisa de que una comprensión más profunda de las emociones humanas requiere un enfoque multidisciplinario (Bankins S et al., 2023), (Budhwar P et al., 2023). Así, la intersección entre la tecnología y la psicología a través del análisis de voz no solo ofrece perspectivas innovadoras, sino que también plantea desafíos éticos y prácticos que deben ser abordados para su correcta implementación (Kuwaiti AA et al., 2023), (Yogesh K Dwivedi et al., 2022). El análisis de voz mediante inteligencia artificial para identificar emociones ha sido objeto de diversas aproximaciones metodológicas que enriquecen la comprensión del fenómeno. Un enfoque destacado ha sido el uso de algoritmos de aprendizaje automático, los cuales permiten la clasificación de emociones a partir de características acústicas.

Estudios como los de (Najjar R, 2023) y (Kamalov F et al., 2023) han demostrado que metodologías que integran redes neuronales profundas ofrecen resultados superiores en comparación con enfoques más tradicionales. Estos autores evidencian que la riqueza de datos y la complejidad de los modelos entrenados son cruciales para lograr una identificación emocional precisa. A pesar de las contribuciones observadas, la revisión ha señalado varias limitaciones críticas en la literatura actual. La mayoría de los estudios se han llevado a cabo en contextos lingüísticos y culturales específicos, lo que limita la generalización de los hallazgos a otras poblaciones (Budhwar P et al., 2023), (Kuwaiti AA et al., 2023). Además, la dependencia de bases de datos que no

representan la diversidad de las expresiones emocionales es un déficit importante que afecta la creación de modelos más robustos y universales de análisis emocional (Yogesh K Dwivedi et al., 2022). Asimismo, la validación de los modelos desarrollados en entornos del mundo real sigue siendo una necesidad no completamente abordada en la literatura (Park S et al., 2022).

Reconocimiento de emociones del habla (Speech emotion recognition)

Definimos un sistema Speech emotion recognition (SER), como un conjunto de metodologías que procesan y clasifican señales de voz para detectar emociones inherentes. Será beneficioso comprender mejor las emociones para optimizar el proceso de clasificación. Existen diversos enfoques para modelar las emociones, y aún es un problema abierto; sin embargo, los modelos discretos y dimensionales se utilizan comúnmente. Un sistema SER requiere un clasificador, un constructo de aprendizaje supervisado, que se entrenará para reconocer emociones en nuevas señales de voz. Un sistema supervisado de este tipo implica la necesidad de datos etiquetados que contengan emociones inherentes. Los datos requieren preprocesamiento antes de poder extraer sus características. Las características son esenciales para un proceso de clasificación, ya que reducen los datos originales a sus características más importantes. Para las señales de voz, se pueden clasificar en cuatro grupos: prosódico, espectral, calidad de voz y características basadas en el operador de energía de Teager. El clasificador puede fortalecerse incorporando características adicionales de otras modalidades, como la visual o la lingüística, según la aplicación y la disponibilidad. Todas estas características se transfieren al sistema de clasificación, que dispone de una amplia gama de clasificadores (Berkehan, 2020).

Transformada Discreta de Fourier (TFD)

Según De la Fraga (2001), Una secuencia periódica puede ser representada por series de Fourier. Con la correcta interpretación, la misma representación puede ser aplicada a secuencias de duración finita. La representación de Fourier resultante para secuencias de duración finita es lo que se conoce como la transformada discreta de Fourier (TDF). Se puede representar una secuencia de duración finita de largo N por una secuencia periódica con periodo N, un periodo de la cual es idéntica a la secuencia de duración finita. Consideremos una secuencia de duración finita x(n) de largo N de forma que x(n) = 0 excepto en el intervalo 0 ≤ n ≤ (N −1). Claramente una secuencia de largo M menor que N también puede considerarse de largo N, teniendo amplitud cero los últimos (N − M) puntos del intervalo. La secuencia periódica correspondiente de periodo N, para la cual x(n) es un periodo, será denotada por $\overline{x}(n)$ y está dada por:

$$\overline{x}(n) = \sum_{r=-\infty}^{\infty} x(n + rN)$$

Dado que x(n) es de largo finito N no hay sobrelapamiento entre los términos x(n + rN) para diferentes valores de r. Así, la ecuación anterior puede ser escrita alternativamente como:

$$\overline{x}(n) = x(x\%N)$$

donde % indica la operación modulo. La secuencia de duración finita x(n) es obtenida a partir de $(\overline{x})$(n) extrayendo un periodo:

$$x(n) = \begin{cases} \overline{x}(n), & 0 \leq n \leq N-1 \\ 0, & de\ otro\ modo \end{cases}$$

Por conveniencia en la notación, es útil definir la secuencia rectangular R_N(n) dada por:

$$R_N(n) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & de\ otro\ modo \end{cases}$$

Con esta notación la ecuación de arriba puede escribirse como:

$$x(n) = \overline{x}(n)R_N(n)$$

Con lo cual, con las series discretas de Fourier, las ecuaciones quedarían de la siguiente forma:

$$X(k) = \begin{cases} \sum_{n=0}^{N-1} x(n)W_N^{kn}, & 0 \leq k \leq N-1 \\ 0, & de\ otro\ modo \end{cases}$$

$$x(n) = \begin{cases} \frac{1}{N}\sum_{n=0}^{N-1} X(k)W_N^{-kn}, & 0 \leq k \leq N-1 \\ 0, & de\ otro\ modo \end{cases}$$

En donde, el par de transformadas dadas por las ecuaciones anteriores se conocen como la Transformada Discreta de Fourier (TFD), en donde, la primera ecuación representa la transformada de análisis y la segunda, la transformada de síntesis.

Transformada Rápida de Fourier (FFT)

La Transformada Rápida de Fourier (FFT) consiste en algunos métodos para calcular la DFT de manera eficiente debido a la importancia que tiene esta en las aplicaciones de tratamiento digital de señales como el filtrado, análisis de la correlación y análisis espectral. Uno de estos métodos consiste en la descomposición de una DFT de N puntos en transformadas DFT sucesivamente más pequeñas. Este método básico nos lleva a una familia de algoritmos de cálculo eficientes conocidos colectivamente como algoritmos FFT, que se puede denotar de la siguiente forma y que correspondería a la función de la ventana de Hamming la cual proporciona más suavizado a través de la

operación de convolución en el dominio de la frecuencia (Jaramillo & Chuquimarca, 2022).

$$x(n) = 0{,}54 - 0{,}46\,cos\left[\frac{2\pi n}{M-1}\right]$$

Enmarcado y enventanado de señales

La idea detrás de dividir señales en "marcos" distintos es descomponer la señal de datos sin procesar en marcos donde la señal tiende a ser más estacionaria. Para obtener características acústicas estables, el habla debe examinarse durante un período de tiempo suficientemente corto. Con respecto a la señal del habla, se informa que el período de 20-30 ms es un segmento cuasi estacionario (QSS), ya que se muestra que el tiempo entre dos cierres glóticos es de alrededor de 20 ms. Sin embargo, se informa que las voces vocálicas se capturan en 40 ms- 80 ms (Yang et al., 2025). Por lo tanto, las mediciones espectrales a corto plazo generalmente se llevan a cabo en ventanas de 20 ms, y cada marco se superpone 10 ms con el siguiente. Las superposiciones de marcos de 10 ms permiten rastrear las características temporales de la señal del habla. Con la superposición de marcos del habla, la representación del sonido estaría aproximadamente centrada en algún marco.

En cada fotograma, se aplica una ventana para estrechar la señal hacia el borde. En general, las ventanas de Hanning y Hamming se encuentran entre las más conocidas. Estas ventanas pueden mejorar los armónicos, suavizar los bordes y disminuir el efecto de borde al aplicar una DFT a la señal. La ilustración 2 describe las ventanas rectangulares de Hamming y Hanning en los dominios del tiempo y la frecuencia.

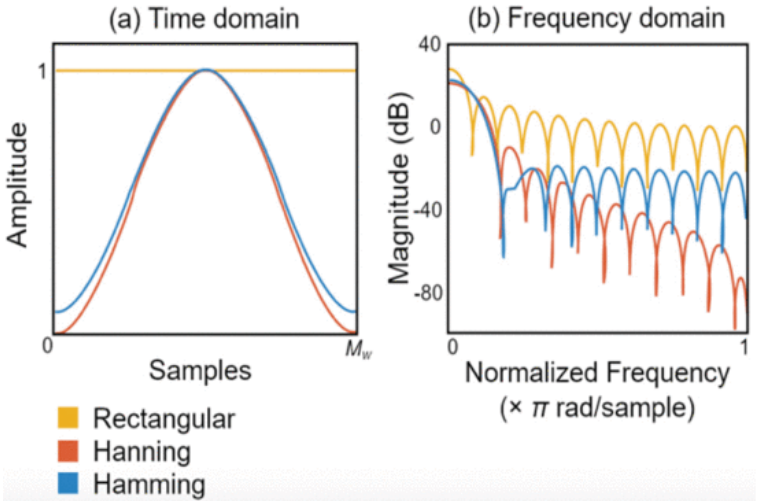


Ilustración 3. Ventanas rectangulares de Hanning y Hamming en los dominios del tiempo y la frecuencia. (Muhamad et al., 2025).

Espectro de frecuencias

Tomás (2015) nos dice que la mayoría de los sonidos están compuestos por varias frecuencias diferentes. El teorema de Fourier afirma que toda señal periódica compleja puede descomponerse en una suma de señales sinusoidales de frecuencias y amplitudes diferentes. Esta descomposición se denomina espectro de frecuencias, y se representa mediante un gráfico con frecuencias en las abscisas y amplitudes en las ordenadas, en el que se visualizan las respectivas amplitudes de todas las frecuencias que componen un sonido.

El espectro cambia constantemente, por lo que las gráficas mencionadas anteriormente representan sólo una porción de sonido que se ha analizado. Esta teoría se aplica de forma rigurosa a las señales totalmente periódicas, pero los sonidos nunca lo son plenamente, pues siempre varían a lo largo del tiempo. El análisis de Fourier se aplica también a señales variables en el tiempo. Una gráfica que represente la variación del espectro a lo largo del tiempo nos da una idea de la evolución de la amplitud de las distintas frecuencias. Esta gráfica puede dibujarse de forma tridimensional, representando los distintos espectros a lo largo del tiempo, compuesta por una sucesión de "rebanadas" temporales, en la que cada una muestra el aspecto del espectro en un instante dado. En la siguiente ilustración, se muestra el espectro de frecuencias de sonido tridimensional. La frecuencia se representa en el eje x, el tiempo en el eje y (aumenta al acercarse hacia el observador), y la amplitud de estas frecuencias en el eje z. Asimismo, en la parte derecha de esta figura, se muestra el espectro de frecuencias o “rebanada” correspondiente al instante inicial t=0.

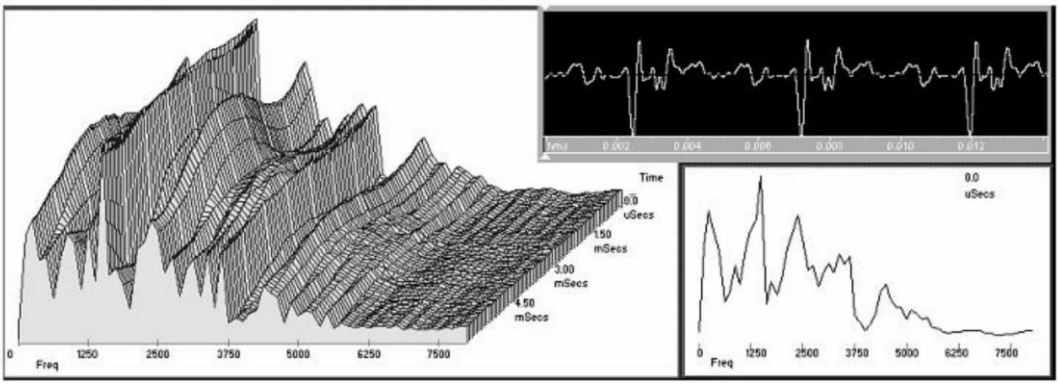


Ilustración 4. Evolución temporal del espectro de frecuencias.

Otra forma de representar esta variación es dibujando un sonograma o espectrograma. Este tipo de gráficas presentan la variación del contenido frecuencial respecto al tiempo, donde el tiempo se presenta en el eje horizontal (o eje x), la frecuencia en el eje vertical (o eje y) y la amplitud se dibuja a través de distintos colores de la traza. Como se puede apreciar en la siguiente ilustración.

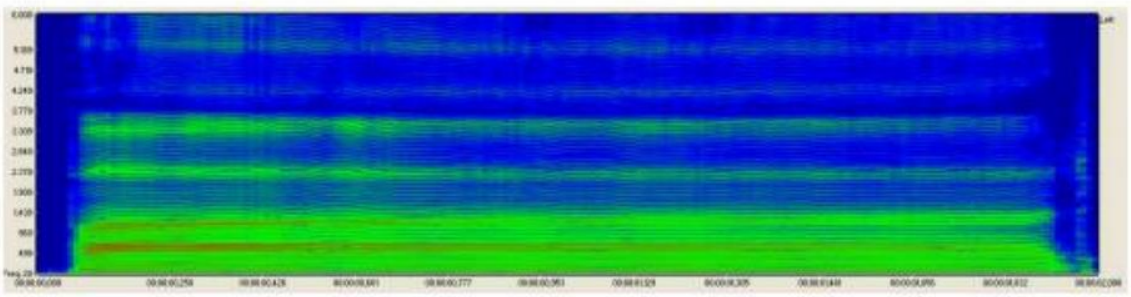


Ilustración 5. Espectograma de un sonido vocálico.

Coeficiente cepstral de frecuencia Mel (MFCC)

La extracción y representación de características tiene un impacto significativo en el rendimiento de cualquier método de aprendizaje automático. El Coeficiente de Cepstrum de Frecuencia de Mel (MFCC) está diseñado para modelar las características de la señal de audio y se utiliza ampliamente en diversos campos (Abdul & Al-Talabani, 2022). Actualmente, existen numerosas técnicas de extracción de características en diversos campos, basadas en los datos brutos. En la mayoría de los campos, la detección de armónicos y bandas laterales de la señal, tanto en el dominio temporal como en el frecuencial, es fundamental para cualquier sistema de reconocimiento de patrones. El espectro de potencia mediante la Transformada Rápida de Fourier (FFT) se utiliza para capturar los armónicos y las bandas laterales de la señal en el dominio temporal. Por otro lado, el cepstrum, como el Coeficiente de Cepstrum de Frecuencia Mel (MFCC) y el Coeficiente de Cepstrum de Tono Gamma (GTCC), permite extraer armónicos y bandas laterales de la versión espectral de la señal (Liang et al., 2013).

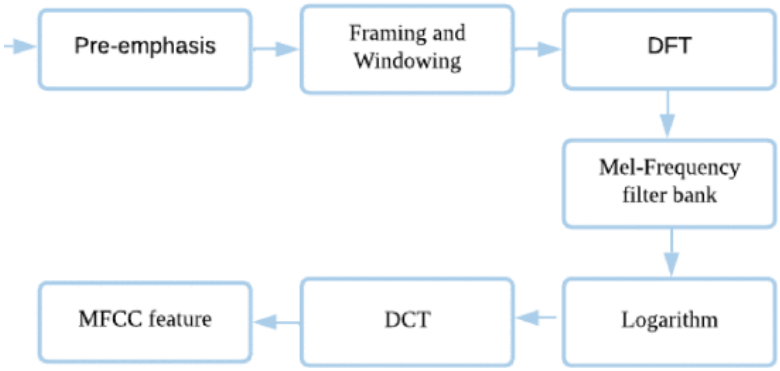


Ilustración 6. Implementación del MFCC. (Abdul & Al-Talabani, 2022)

Los Coeficientes Cepstrales en la Escala de Mel (MFCC) representan la amplitud del espectro del habla de manera compacta, esto los ha vuelto la técnica de extracción de características más usada en reconocimiento del habla. En la Ilustración 4, se muestra

el proceso para la elaboración de un vector característico de MFCC. Primeramente, se aplica un filtro de pre-énfasis a la señal y posteriormente se divide la misma en tramas y se le aplica una función de ventaneo, en este caso una ventana de Hamming de 20 ms. El ventaneo sirve para eliminar los bordes de la señal y darle una acentuación a la parte central de la trama para su análisis. Al obtener la Transformada Discreta de Fourier (DFT) de cada trama se utiliza la amplitud del espectro, y esta información es pasada al dominio de Mel mediante el Banco de Filtros. La escala Mel se basa en mapear entre la frecuencia actual al pitch que percibe, un escucha humano simulado, esta escala es lineal por debajo de 1 kHz y logarítmica por encima de este umbral. Después se obtiene el logaritmo de la señal y finalmente se aplica la Transformada de Coseno Discreta (DCT), de este vector obtenido se toman la cantidad de coeficientes deseados por trama (Martínez Mascorro & Aguilar Torres, 2013).

Redes Neuronales Recurrentes (RNN)

Las Redes Neuronales Recurrentes (RNN) se han consolidado como una herramienta fundamental en el ámbito del aprendizaje automático, especialmente en el procesamiento de datos secuenciales. Estos modelos conexionistas son capaces de modelar la dependencia temporal de las series de datos, lo que permite analizar patrones complejos a lo largo del tiempo. Las redes neuronales recurrentes (RNN) son modelos conexionistas que capturan la dinámica de las secuencias mediante ciclos en la red de nodos. Esta propiedad las hace especialmente efectivas en tareas como la traducción automática y la generación de texto, donde la relación entre elementos distantes en la secuencia es crítica. Además, innovaciones como el uso del GCRN, que combina RNN con redes convolucionales, han demostrado potenciar la precisión y velocidad de aprendizaje al integrar información espacial y dinámica en diferentes contextos, como la detección de cambios en imágenes multitemporales (Mou et al., 2018). Así, las RNN continúan evolucionando, facilitando soluciones a problemas complejos en distintos dominios.

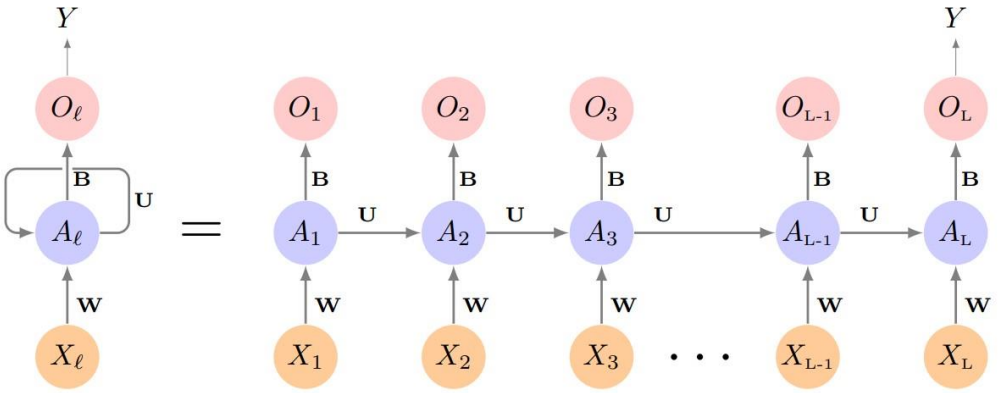


Ilustración 7. Redes Neuronales Recurrentes. (Amat & Carazo, 2025).

Casos de Uso de las RNN en el Procesamiento del Lenguaje Natural

El procesamiento del lenguaje natural (PLN) ha sido revolucionado por el avance de las redes neuronales recurrentes (RNNs), que han demostrado ser especialmente efectivas para lidiar con secuencias de texto y datos lingüísticos. Entre sus casos de uso más notables se encuentran la generación automática de texto, la traducción de idiomas y el análisis de sentimientos, donde las RNNs pueden captar relaciones temporales y de contexto en las palabras. Por ejemplo, en la identificación de eventos adversos en la atención médica, las RNNs permiten analizar informes clínicos redactados en lenguaje natural para detectar y clasificar incidentes relacionados con errores médicos, mejorando así la seguridad del paciente (Cohan et al., 2017). Asimismo, en el ámbito de la detección de olores en el código fuente, las RNNs, junto con otras arquitecturas de redes neuronales, han mostrado un potencial prometedor para abordar la calidad del software, explorando nuevas fronteras en la intersección entre el PLN y la ingeniería de software (Sharma et al., 2019). Estos ejemplos subrayan cómo las RNNs no solo son herramientas versátiles, sino que también contribuyen significativamente a la resolución de problemas complejos en diversas disciplinas.

Redes Neuronales Convolucionales (CNN)

Las Redes Neuronales Convolucionales (CNN) han revolucionado el campo del aprendizaje automático, especialmente en el ámbito del procesamiento de imágenes y el análisis de datos visuales. Estas arquitecturas son reconocidas por su capacidad para realizar extracciones de características de forma jerárquica, lo que les permite identificar patrones y detalles complejos en imágenes. Este potencial ha llevado a su adopción en diversas aplicaciones, desde la clasificación de imágenes hasta la detección de enfermedades a través de técnicas de imagen médica. Por ejemplo, en proyectos relacionados con el diagnóstico de Alzheimer, se ha demostrado que las CNN pueden ofrecer resultados prometedores al analizar resonancias magnéticas, alcanzando una precisión de hasta el 81.10% en tareas de validación (Lee & Keegan, 2023). Además, su influencia en el desarrollo de nuevas tecnologías médicas resalta la importancia de las CNN en la intersección entre la inteligencia artificial y la salud, lo que subraya su relevancia en la investigación actual (Waheed Javed & Parveen, 2023).

I

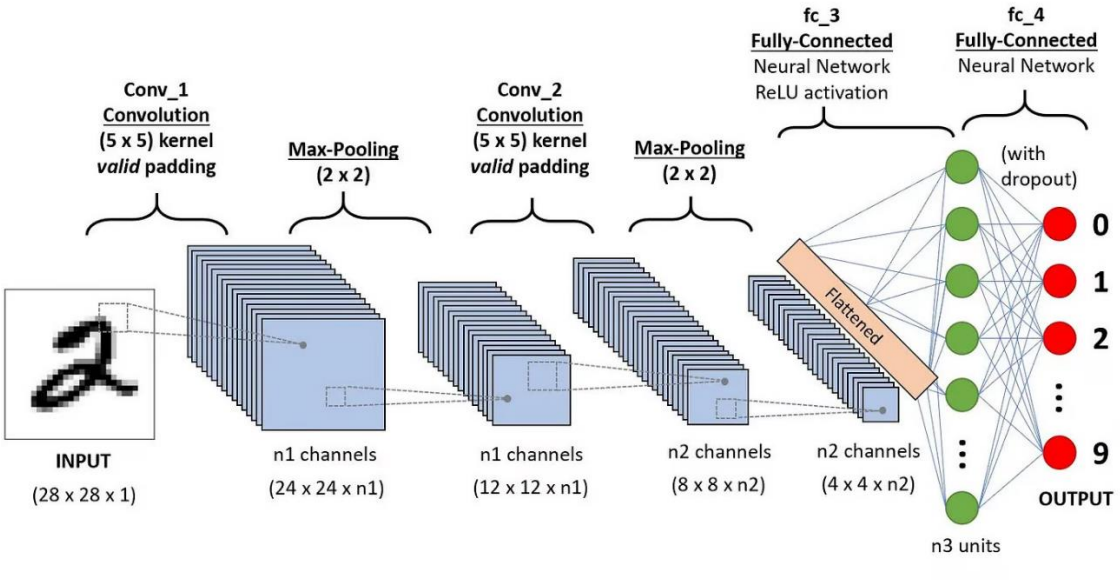


Ilustración 8. Esquema de la red CNN. (datacamp, 2024)

Estructura y Funcionalidad de las CNN

La estructura y funcionalidad de las Redes Neuronales Convolucionales (CNN) son fundamentales para comprender su eficacia en diversas aplicaciones, desde la visión por computadora hasta el procesamiento de datos médicos. Estas redes están diseñadas para procesar datos estructurados en forma de cuadrículas, como imágenes, utilizando capas convolucionales que extraen características jerárquicas y patrones espaciales específicos. El enfoque multitarea que integran las CNN, junto con la utilización de arquitecturas avanzadas, permite realizar análisis complejos, como la detección de plagio en código o la clasificación de enfermedades a partir de imágenes médicas. Por ejemplo, un sistema que combina CNN y LSTM ha demostrado ser efectivo en la detección de diversas formas de plagio en C, logrando altas tasas de precisión y record a través de la extracción multietapa de características (Surendran, 2024). Además, la capacidad de las CNN para trabajar con datos clínicos y de imagen mejora la toma de decisiones en el ámbito de la salud, optimizando así los diagnósticos y tratamientos (Desai, 2020).

Redes de Memoria a Largo y Corto Plazo (LSTM)

Al abordar las Redes de Memoria a Largo y Corto Plazo (LSTM) dentro del contexto de las redes neuronales, es crucial entender su papel en la resolución de problemas complejos relacionados con secuencias temporales y datos secuenciales. Estas redes son altamente efectivas para gestionar dependencias a largo plazo, lo que las convierte en una opción invaluable en diversas aplicaciones como el reconocimiento de acciones en video y la predicción de series temporales. Por ejemplo, las redes

neuronales recurrentes de memoria a corto y largo plazo (LSTM) han demostrado ser herramientas eficaces para la predicción de series temporales complejas, como los precios de acciones en mercados financieros. Además, estudios recientes han evidenciado que la combinación de LSTM con redes convolucionales mejora notablemente la precisión en el reconocimiento de acciones, mostrando resultados superiores al integrar características de ambos tipos de redes en un marco de fusión (Chopra et al., 2023), (Zhao et al., 2017).

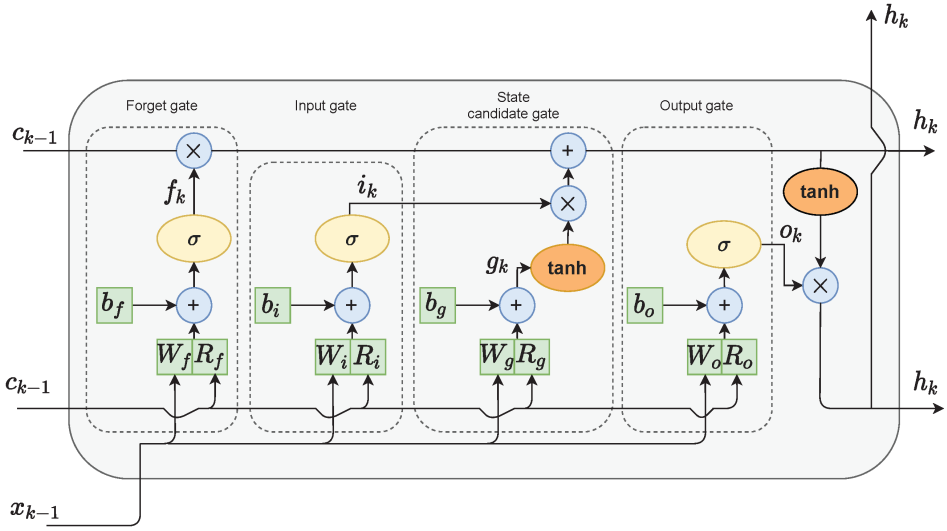


Ilustración 9. Estructura de Red LSTM. (Zarzycki & Ławryńczuk, 2021).

Implementación de las LSTM en la Predicción de Series Temporales

La implementación de redes neuronales de tipo LSTM en la predicción de series temporales ha emergido como una metodología prominente debido a su capacidad para capturar dependencias a largo plazo en los datos. A diferencia de otras arquitecturas, como las redes neuronales convolucionales (CNN) o las redes recurrentes tradicionales (RNN), las LSTM están diseñadas específicamente para recordar información relevante a través de secuencias extensas, lo cual es crucial en aplicaciones como la predicción financiera o meteorológica. La integración de características de memoria en estas redes permite que manejen problemas comunes, como el desvanecimiento de gradientes, lo que resulta en un rendimiento superior e incrementa la precisión de las predicciones. Estudios recientes han demostrado que los sistemas que combinan LSTM con otras técnicas avanzadas logran mejorar significativamente las tasas de reconocimiento en conjuntos de datos estándar (Zhao et al., 2017), reflejando su idoneidad en el ámbito de análisis de datos complejos (Tay et al., 2017).

Red neuronal GRU

La red GRU es una modificación del concepto LSTM, cuyo objetivo es reducir el coste computacional de la red. Existen algunas diferencias entre las arquitecturas, principalmente:

1. La celda GRU carece de puerta de salida, por lo tanto, tiene menos parámetros;
2. Se descarta el uso del estado de la celda. El estado oculto sirve como memoria de trabajo y de largo plazo de la red.

La disposición de la celda GRU única se presenta en la siguiente ilustración.

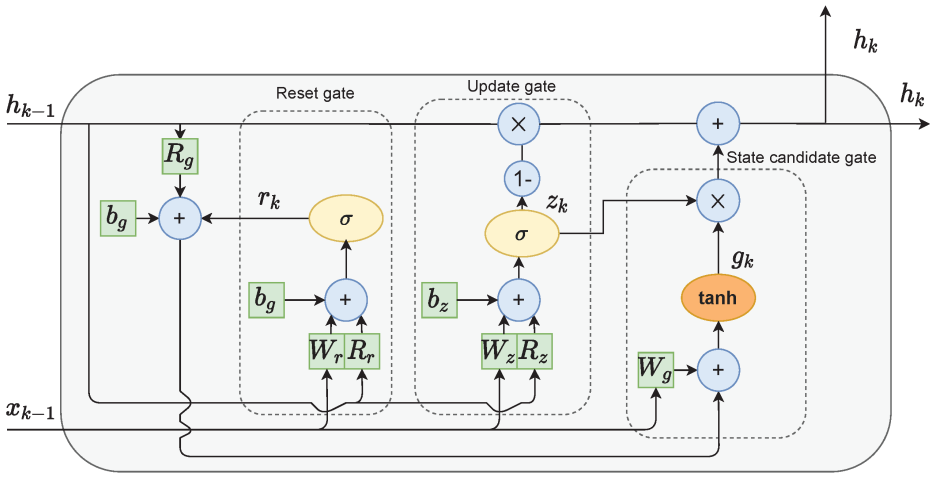


Ilustración 10. Estructura de red GRU.(Zarzycki & Ławryńczuk, 2021).

Visión General de las Arquitecturas CNN, RNN y LSTM

En el ámbito de las redes neuronales profundas, las arquitecturas CNN, RNN y LSTM han revolucionado el campo del aprendizaje automático, cada una con funcionalidades específicas que abordan diferentes tipos de datos y tareas. Las redes neuronales convolucionales (CNN) son especialmente efectivas en el procesamiento de datos espaciales, como imágenes, al extraer características jerárquicas a través de sus capas convolucionales. Por otro lado, las redes neuronales recurrentes (RNN), como su nombre indica, están diseñadas para manejar datos secuenciales, lo que las hace ideales para tareas relacionadas con el lenguaje o el reconocimiento de patrones temporales. En este contexto, las LSTM, que son una variante de RNN, ofrecen una ventaja notable al gestionar dependencias a largo plazo en las secuencias, permitiendo una comprensión más profunda de la dinámica temporal. Como se ha señalado, las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN) son los dos principales tipos de arquitecturas de redes neuronales profundas ampliamente exploradas para manejar diversas tareas de procesamiento de

lenguaje natural. Además, la integración de estas arquitecturas, como se menciona en estudios recientes, demuestra su capacidad para realizar análisis contextuales complejos en aplicaciones innovadoras, como la detección de cambios en imágenes multiespectrales (Mou et al., 2018).

4. Desarrollo del proyecto y resultados

4.1. Metodología

Para el presente estudio se propuso desarrollar el siguiente esquema de trabajo de acuerdo a lo indicado en la ilustración 11.

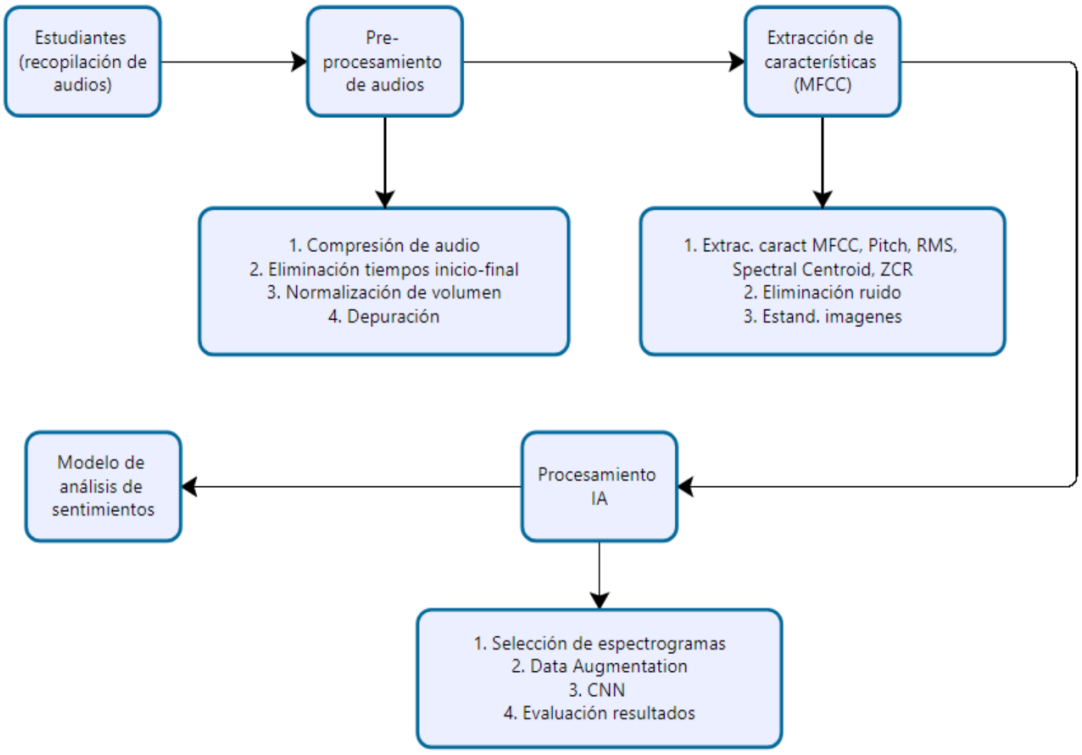


Ilustración 11. Propuesta a desarrollar. Elaboración propia

El presente trabajo es de tipo experimental, con enfoque cuantitativo y de tipo aplicado, basado en análisis audios recolectados de los estudiantes para evaluar los espectogramas o imágenes de las señales de audio y así desarrollar un modelo que corresponde a una solución en inteligencia artificial como el uso de redes neuronales convolucionales para el reconocimiento de emociones.

Población y muestra.

La población estuvo conformada por todos los estudiantes ingresantes a la universidad pública del presente estudio, en la ciudad de Lima-Perú. La muestra estuvo conformada por todos los audios que se recolectaron por los estudiantes y que fueron en su totalidad 4,720 audios divididos en seis estados emocionales como son: alegría, enojo, miedo, neutral, sorpresa y tristeza. Estos audios fueron recolectados a través de la lista indicada en la tabla x, que se presenta a continuación.

EMOCION	TEXTO
Alegría-1	¡Acabo de recibir la mejor noticia de mi vida!
Alegría-2	Hoy es mi cumpleaños y me siento increíblemente amado.
Alegría-3	¡Conseguí el trabajo que tanto deseaba!
Alegría-4	Mis amigos me prepararon una fiesta sorpresa, ¡estoy emocionado!
Alegría-5	El bebé dio sus primeros pasos hoy, ¡qué alegría!
Alegría-6	Finalmente me gradué después de años de esfuerzo.
Alegría-7	¡El equipo ganó en el último minuto, qué emoción!
Alegría-8	Recibí un abrazo inesperado que me llenó de felicidad.
Alegría-9	Hoy el sol brilla y todo parece perfecto.
Alegría-10	¡Mi perro corrió a saludarme como siempre, me encanta!
Tristeza-1	Nadie vino a mi fiesta... me siento tan solo.
Tristeza-2	Perdí a mi mascota hoy, el dolor es inmenso.
Tristeza-3	Me despidieron sin explicación alguna.
Tristeza-4	La lluvia refleja exactamente cómo me siento por dentro.
Tristeza-5	Extraño tanto a mi abuela que ya no está aquí.
Tristeza-6	Me dejó después de 10 años juntos.
Tristeza-7	Olvidaron mi cumpleaños otra vez...
Tristeza-8	El médico dio malas noticias sobre los resultados.
Tristeza-9	Me fallaron cuando más los necesitaba.
Tristeza-10	Todo parece salir mal últimamente.
Enojo-1	¡Rompió mi teléfono a propósito!
Enojo-2	Me estafaron y no puedo hacer nada al respecto.
Enojo-3	Nunca llegan a tiempo, ¡es exasperante!
Enojo-4	¿Cómo te atreves a mentirme así?
Enojo-5	El tráfico hoy me hizo perder la paciencia.
Enojo-6	Me cancelaron el vuelo sin compensación.
Enojo-7	¡Deja de criticar todo lo que hago!

Enojo-8	Me hirieron con sus comentarios malintencionados.
Enojo-9	No soporto la injusticia de esta situación.
Enojo-10	Otra vez lo mismo, ¡estoy harto!
Miedo-1	Escuché un ruido en la casa y estoy solo...
Miedo-2	El avión empezó a sacudirse violentamente.
Miedo-3	Creo que alguien me sigue por la calle.
Miedo-4	Tengo que hablar en público y me aterra.
Miedo-5	Los resultados médicos podrían cambiar mi vida.
Miedo-6	El perro corrió hacia mí gruñendo.
Miedo-7	La película de terror me dejó sin dormir.
Miedo-8	No sé nadar y caí al agua profunda.
Miedo-9	El sonido de los truenos me paraliza.
Miedo-10	Me perdí en el bosque al anochecer.
Sorpresa-1	¡Me propusieron matrimonio en plena cena!
Sorpresa-2	¿En serio gané este premio tan valioso?
Sorpresa-3	¡No esperaba verte aquí después de años!
Sorpresa-4	El final de la película me dejó boquiabierto.
Sorpresa-5	Mi jefe me dio un aumento inesperado hoy.
Sorpresa-6	¡Mi equipo perdió contra el último lugar!
Sorpresa-7	Descubrí que mi hermano gemelo existía a los 30.
Sorpresa-8	¡El pastel tenía un anillo dentro!
Sorpresa-9	Me devolvieron la billetera con todo el dinero.
Sorpresa-10	¡El test de embarazo salió positivo!
Neutral-1	El informe debe entregarse antes del viernes.
Neutral-2	Compré pan y leche en el supermercado.
Neutral-3	La reunión empieza a las 3:00 pm puntual.
Neutral-4	El clima hoy estará parcialmente nublado.
Neutral-5	El autobús pasa cada 15 minutos.
Neutral-6	Mi dirección es Calle Principal #123.
Neutral-7	La computadora necesita actualización.
Neutral-8	El libro tiene 320 páginas en total.
Neutral-9	El vuelo despegará a las 8:00 am.
Neutral-10	La receta lleva harina, huevos y leche.

Tabla 1. Listado de emociones recolectadas a través de audios. Elaboración propia

Para recolectar la información, se desarrolló un formulario en Google, con el propósito de que puedan ser cargados a un repositorio central, del mismo modo, se desarrolló un manual de procedimiento para los estudiantes, de tal forma, que puedan recibir los pasos detallados adecuadamente. Adicionalmente, se realizó la exposición de como se deberían grabar los audios.

Se utilizó un software que permitió grabar cada uno de los audios a entregar, el cual no necesitaba ser instalado en la computadora, del mismo modo, permitía configurar audio de alta calidad al momento de grabar, en formato estéreo y en 48000 Hz, que posteriormente, serán convertidos a monoaural con 16000 Hz. Todos los audios fueron grabados en formato sin compresión (formato *.wav), aquellos que fueron grabados en otros formatos no fueron considerados para el presente estudio.

Posteriormente, se procedió con el preprocesamiento de los audios recolectados, en donde se realizó una compresión del audio con el objetivo de evitar picos innecesarios, luego se hizo la eliminación de los espacios no audibles al inicio y al final de cada audio, se identificó además que muchos audios se encontraban en bajo volumen, por lo cual se hizo un reprocesamiento, sin ganar ni perder calidad de los mismos, con el fin de normalizar el volumen correspondiente, adicional a esto, se utilizó un proceso que segmentaba el audio en sub bandas y eliminaba la banda más débil, con el propósito de eliminar el ruido y de esta forma, maximizar la voz humana, finalmente, se evaluó y estandarizó el tamaño de los audios para el siguiente proceso.

Con la limpieza y estandarización de audios realizada, se procedió a realizar la extracción de características, para lo cual, se utilizó MFCC o Coeficientes Cepstrales de Frecuencia Mel, que son una representación compacta de las características espectrales de cada audio, esto se realizará a través de una librería de Python que permite extraer 40 coeficientes MFCC por ventana de 25 ms con salto de 10 ms, esta representación es la permite identificar las escalas y las características del audio en datos a través de un ventaneo por cada segmento de espacio de tiempo, y que permite identificar y descomponer el comportamiento de la señal de audio. Adicional a ello, se realizó también la identificación de otros valores específicos sobre el conjunto de datos, tales como el Pitch o la intensidad o percepción de la altura del tono, RMS, que es una medida de potencia promedio, el centroide espectral, que indica el centro de espectro de la frecuencia, ZCR o tasa de cruce por cero, que indica cuántas veces ha cambiado de signo la frecuencia, así como la disminución de ruido y la estandarización de imágenes por cada clase.

Desarrollados los espectogramas, a través de los coeficientes de Mel, (MFCC), se procedió a desarrollar el modelo que permitiera identificar los diversos estados de emociones recolectados para este propósito, el cual, posteriormente, fue evaluado con otros modelos para identificar la precisión del mismo.

Para completar el propósito del estudio a realizar, fue necesario revisar adecuadamente los parámetros correspondientes al reconocimiento de emociones del habla peruana, principalmente del habla en Lima, donde se identificaron parámetros

diferenciados con respecto del estándar de uso en la región, e incluso un ligero diferencial con la región andina, de tal forma, que el modelo se puede ajustar mejor al conjunto de audios recogidos.

	Pitch (Hz)	Energy (dB)	Spectral Centroid (Hz)	ZCR	Weight
Alegria	(220, 300)	(-18, -8)	(1500, 2500)	(0.08, 0.15)	1.2
Enojo	(250, 350)	(-15, -5)	(2200, 3200)	(0.15, 0.25)	1.1
Miedo	(170, 260)	(-23, -13)	(2900, 3900)	(0.11, 0.19)	1.0
Tristeza	(90, 140)	(-28, -23)	(700, 1400)	(0.04, 0.09)	1.0
Sorpresa	(210, 330)	(-20, -10)	(1900, 2900)	(0.09, 0.17)	1.0
Neutral	(140, 190)	(-23, -16)	(900, 1400)	(0.06, 0.1)	1.0

Tabla 2. parámetros aproximados del habla en la ciudad de Lima. Elaboración propia

4.2. Planteamiento del problema

El propósito de la presente investigación radica en explorar las capacidades de la inteligencia artificial (IA) en el análisis de la voz como herramienta para identificar emociones, un área que ha cobrado relevancia en diversas disciplinas, desde la psicología hasta el desarrollo de tecnologías de comunicación e información. Este trabajo no solo presenta un marco teórico sobre las metodologías empleadas en el análisis de la voz, sino también discutir sus implicaciones prácticas y éticas. En este sentido, es fundamental considerar los desafíos que surgen en la implementación de estas tecnologías, como se evidenció en el proyecto Critically Exploring Biometric AI Futures, el cual abordó las preocupaciones sociales, éticas y legales relacionadas con el uso de IA en entornos sensibles como el de la ley y la seguridad (Connon et al., 2023).

Es por esto que el presente trabajo de investigación se encuentra en la búsqueda de hacer reconocimiento de voz utilizando inteligencia artificial para determinar los estados emocionales de los estudiantes ingresantes a la universidad, de esta forma, identificar la percepción de los estudiantes en sus emociones, debido a que los estudiantes del primer año de estudios no se les realiza un estudio inicial adecuado, ni seguimiento sobre su estado emocional, lo cual es impredecible en eventos futuros que se pueden desencadenar dentro de la institución universitaria tales como agresión, fobia, suicidio, falta de valoración de la vida, entre otros durante sus clases presenciales por el posible estrés o los cursos que se llevan en la etapa académica.

PROBLEMA PRINCIPAL

¿De qué manera el análisis de voz con inteligencia artificial puede identificar el estado emocional de los estudiantes en una universidad?

JUSTIFICACIÓN E IMPORTANCIA.

En el contexto educativo actual en el Perú, el componente emocional de los estudiantes se ha consolidado como un factor determinante en el rendimiento académico, la motivación y la calidad del proceso enseñanza-aprendizaje. Las emociones influyen directamente en la atención, la retención de información y la participación activa en el aula, convirtiéndose en un indicador clave para comprender el bienestar y la disposición del estudiante para aprender.

A pesar de su relevancia, las emociones suelen ser abordadas de forma subjetiva o poco sistematizada dentro del entorno educativo. En este sentido, la incorporación de tecnologías emergentes como la inteligencia artificial (IA) ofrece una oportunidad innovadora para el monitoreo y análisis automatizado del estado emocional de los estudiantes, permitiendo generar información objetiva y en tiempo real que sirva de insumo para mejorar las estrategias pedagógicas.

Como planteamiento de problema fundamental, lo que se busca es poder identificar patrones recurrentes dentro de los espectros de audio de la voz de los estudiantes en los diversos estados emocionales como alegría, miedo, tristeza, neutral, enojo y sorpresa; de tal forma, que se pueda establecer un modelo en inteligencia artificial a través de redes neuronales convolucionales, con el propósito de generar un modelo que permita a partir de un audio recolectado, pueda identificar qué tipo de estado emocional se encuentra en el estudiante, de esta forma, se busca prevenir acciones futuras no deseadas por él mismo.

4.3. Desarrollo del proyecto

A continuación, se describen las diversas etapas desarrolladas en la presente investigación.

Los audios fueron recolectados a través de un formulario en Google, con un instructivo muy adecuado a los participantes, donde se indicaba en forma detallada, cuál sería el procedimiento a a realizar para recolectar los audios solicitados. La ilustración 12 muestra parte del instructivo que fue usado para la recolección.

Actividad: Reconocimiento de Emociones

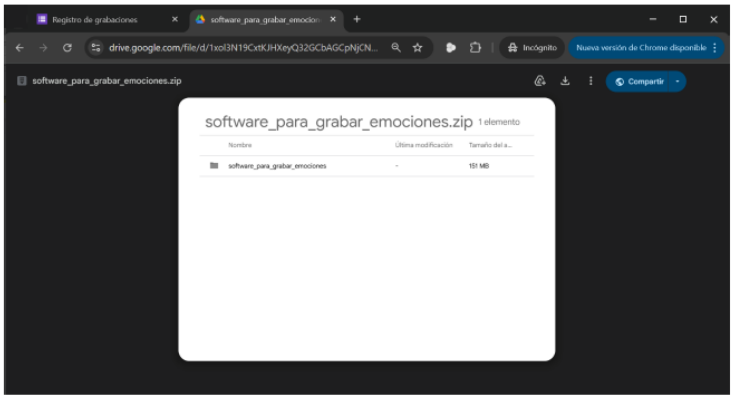
Manual para la grabación y envío de archivos de emociones

Estimado estudiante, saludos cordiales, a continuación, te alcanzo los pasos a seguir para que puedas grabar las emociones solicitadas con el procedimiento solicitado en este documento.

Como primer paso, deberás descargar los archivos de grabación del siguiente enlace:

https://drive.google.com/file/d/1xol3N19CxtKJHXeyQ32GCbAGCpNjCNnl/view?usp=sharing

Aparecerá una ventana como la indicada aquí abajo donde podrás descargar la carpeta que se encuentra zipeada y que se indica en la imagen, recuerda que puedes descargar el archivo sin problemas ya que han sido verificados y no contienen ningún virus.



Si no visualizas directamente esta ventana, posiblemente debas ingresar primero a tu cuenta de Google para poder visualizarla.

Luego de descargarlo en la carpeta de tu preferencia, deberás descomprimir el archivo y deberás visualizar el siguiente contenido:

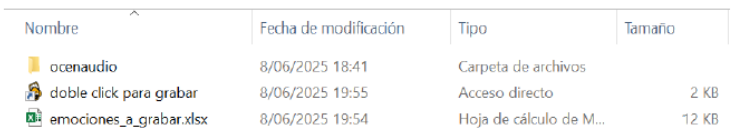


Ilustración 12. Instructivo para grabar y recoger los audios de emociones. Elaboración propia

Del mismo modo, se utilizó un programa que no necesitaba instalarse en la computadora para realizar el proceso de grabación de los audios a enviar, este programa permitió grabar los audios en alta calidad para tener así un mejor conjunto de datos, la ilustración 13 muestra las condiciones que permitía grabar los audios, lo cual permitieron una buena calidad de datos para el estudio. Estas condiciones se indicaron en el manual de grabación para la recolección de información.

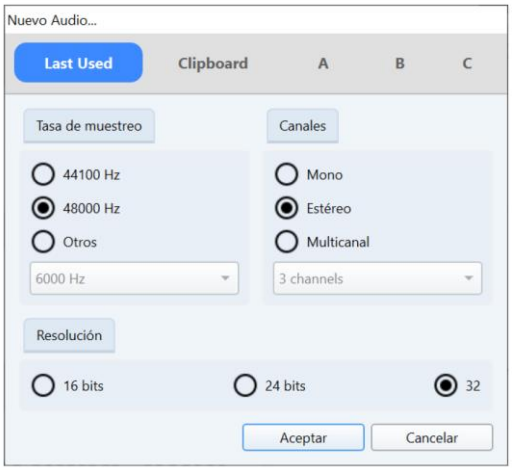


Ilustración 13. Calidad de los audios durante la grabación. Elaboración propia

El proceso de recolección de audios se realizó durante los meses de mayo, junio y julio, permitiendo indicar las condiciones de grabación a todos los participantes, así como indicándoles la confidencialidad de los mismos, para mantener en todo momento, las condiciones de ética correspondientes en el presente estudio.

Una vez recolectados los audios, se procedió con la compresión correspondiente por cada audio, de tal forma, que los picos innecesarios se eliminan o se suavizan para un adecuado tratamiento. La ilustración 14 muestra la compresión de un audio del estudio para este proceso.

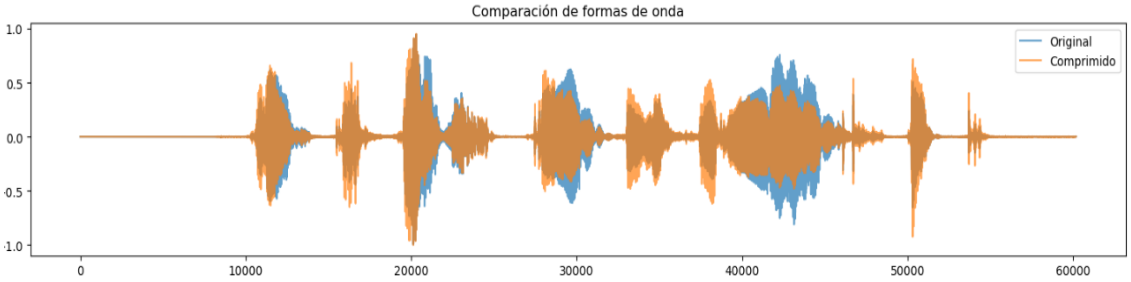


Ilustración 14. Compresión de un audio del dataset del estudio. Elaboración propia

Después de este proceso, se realizó la eliminación de tiempos al inicio y al final de cada audio, para lo cual, se utilizó la librería librosa de Python, con la función librosa.effects.trim(), que permitió recortar los extremos no audibles de cada elemento del dataset.

La normalización del volumen estuvo relacionada con evaluar si un audio se encontraba con un mínimo de decibeles y no se podía escuchar, de tal forma, que se aplicó una ecuación de ganancia sobre el mínimo, hasta un tope correspondiente de acuerdo a la siguiente ecuación:

$$F(x) = 20 * log10(0.9)/x)$$

En donde, a mayor valor de la variable "x", menor sería el incremento de audio. Se utilizó el valor de 0.9 para evitar clipping (límite máximo antes de que pase tener distorsión), de tal forma, que se pudo controlar el valor a aumentar.

Para la extracción de características se utilizó MFCC en descomposición de 13 características, puesto que se estuvo realizando pruebas con 20, 40 y 50, pero los resultados eran los mismos, o incluso, se ajustaban mejor con la descomposición a 13 características de la señal de audio. La ilustración 15 muestra una representación de la descomposición de los datos del estudio.

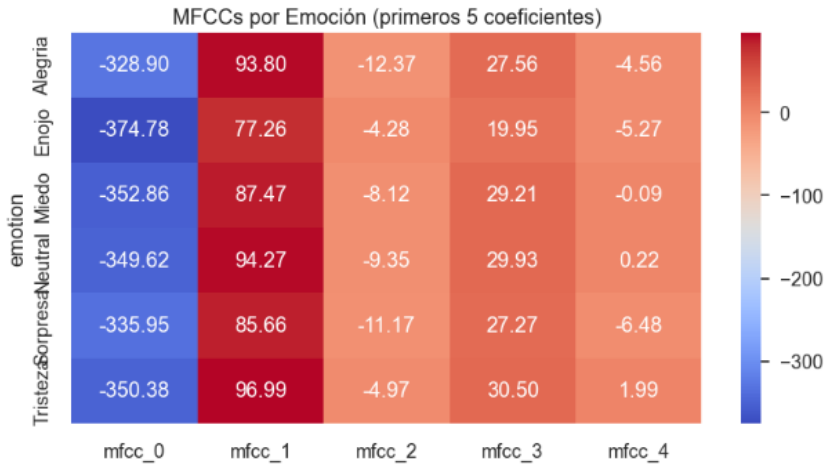


Ilustración 15. Descomposición MFCC. Elaboración propia

Como se puede apreciar, la descomposición de Mel permite pasar a datos el conjunto de audios. Cada uno de los audios se puede representar en sus 13 características principales, a través de un ventaneo por intervalos de milisegundos, (tal como se describió anteriormente), para representar todas las frecuencias de audio. Así mismo, la ilustración 16 muestra la correlación entre las primeras 15 características de los audios del estudio.

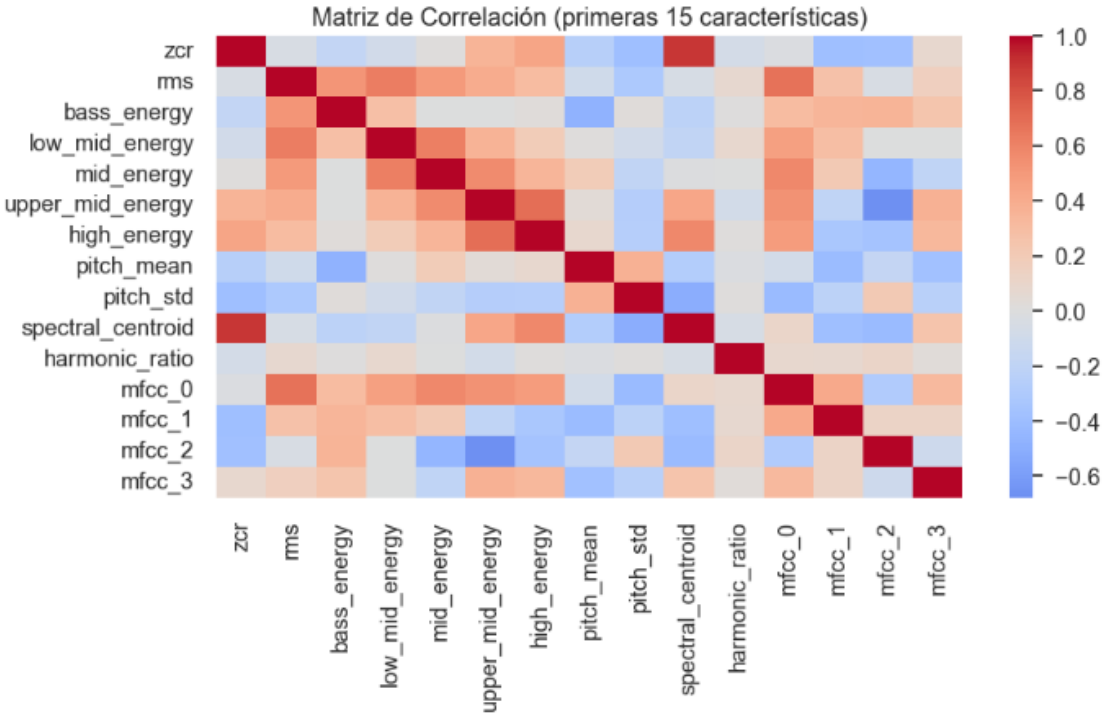


Ilustración 16. correlación de características. Elaboración propia

La diagonal principal muestra el valor de correlación de cada variable consigo misma, por lo tanto, todos los valores en esta diagonal son 1.0 (correlación perfecta, representado en rojo oscuro). Rojo oscuro fuerte representa una correlación positiva fuerte, cuando una variable aumenta, la otra también. Azul oscuro representa una correlación negativa fuerte, cuando una variable aumenta, la otra disminuye. Tonos claros representan una baja o nula correlación.

Dentro de los parámetros acústicos de energía se tiene el rms, que representa la raíz cuadrada media de amplitud, bass_energy, low_mid_energy, mid_energy, upper_mid_energy, high_energy son energía en subbandas específicas del espectro de frecuencia.

Sobre las características prosódicas se tiene el pitch_mean, que representa la frecuencia fundamental media, pitch_std: desviación estándar del pitch y zcr que es la tasa de cruce por cero, indicador de contenido de alta frecuencia.

Las características espectrales y armónicas son el spectral_centroid, que viene a ser el centroide espectral e indica el "brillo" del sonido; la harmonic_ratio: relación armónica, vinculada a la periodicidad de la señal.

Los coeficientes cepstrales de Mel son los indicados como mfcc_0, mfcc_1, mfcc_2 y ,fcc_3 que capturan el envolvente del espectro en bandas perceptuales humanas.

Sobre el pitch, que es la intensidad con la cual se puede identificar un audio, en la ilustración 17, se identifica el pitch promedio para cada una de las emociones recogidas.

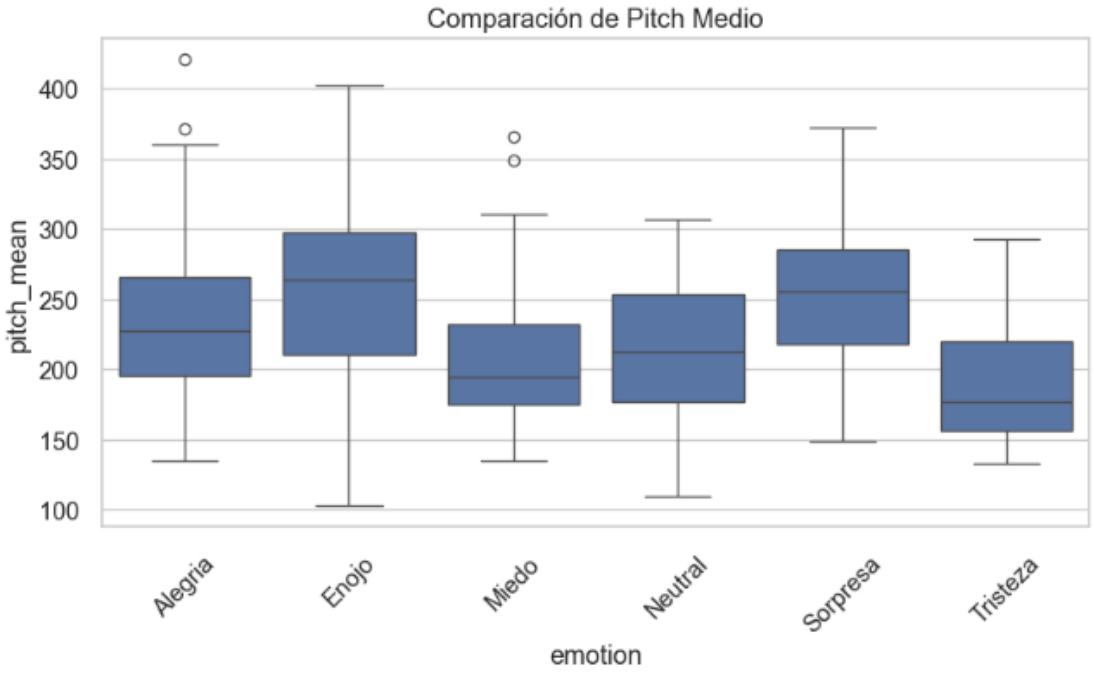


Ilustración 17. Pitch promedio de audios por emoción. Elaboración propia

En donde se puede apreciar los rangos promedios por cada intensidad de emoción, la media, así como algunos valores atípicos dentro de los rangos principales.

Identificando en forma visual, sobre los espectogramas obtenidos, podemos describir los siguientes para algunos audios del estudio.

La ilustración 18 muestra un espectograma de un audio con silencio en los primeros 2 segundos antes del inicio del pitch o del habla del participante. Como se puede apreciar, no existe valores o tonalidades en los primeros segundos, sin embargo, del segundo al cuarto, existe una diferenciación de tonalidades mientras se encontraba hablando. Para este tipo de casos, se eliminaron los 2 primeros segundos debido a que no aportaban nada al presente estudio. Para este caso, el espectograma corresponde a un audio de enojo.

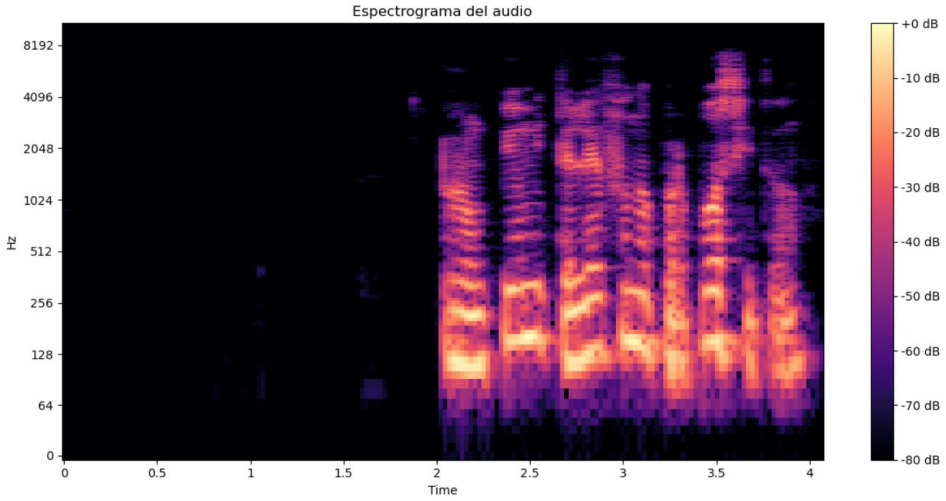


Ilustración 18. Espectograma sin audio inicial. Elaboración propia

La ilustración 19 muestra ahora cuando un audio del estudio se encuentra con ruido, nótese la trama o el entramado existente en todo el espectograma. Aquí se realizaron dos procesos de limpieza de datos, el primero, fue el recorte del primer 0.5 segundos, debido a que al parecer no estaba activo aún el micrófono, pero ya estaba activa la grabación correspondiente. La segunda limpieza corresponde a la eliminación del ruido, identificando sub bandas en el audio y recortando las bandas de menor implicancia sobre los patrones audibles de voz del participante. De forma similar, para la ilustración descrita, esta corresponde a un sentimiento de alegría.

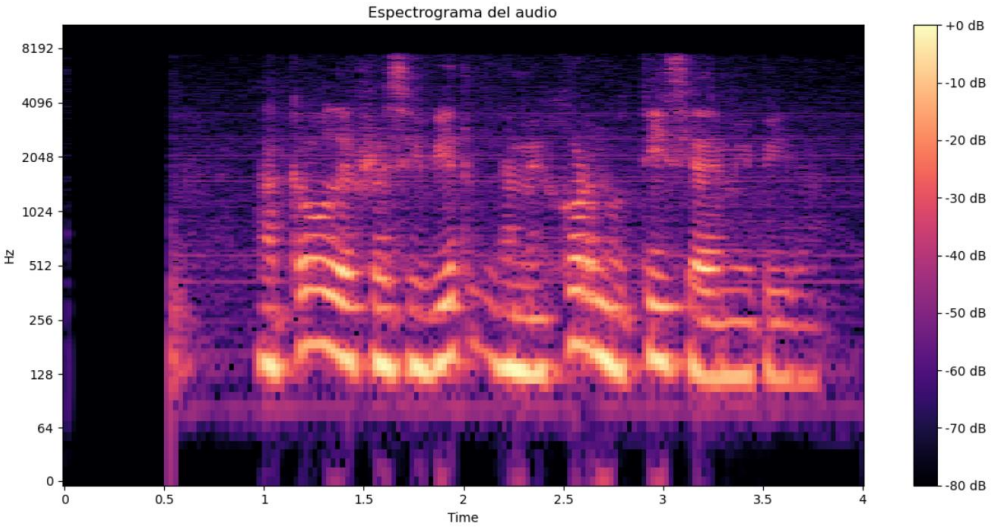


Ilustración 19. Espectograma con ruido. Elaboración propia

La siguiente ilustración, (ilustración 20), muestra un espectograma de un audio de tristeza, en el cual, para la limpieza de datos, también se aplicó un recorte en el inicio y el final de aproximadamente de medio segundo. Nótese la diferencia de tonalidades o de fuerza de voz que experimenta este espectograma con respecto de los dos anteriores, en este caso, las tonalidades son mucho más bajas o tenues, identificándose una fuerza de voz débil frente a los casos presentados anteriormente.

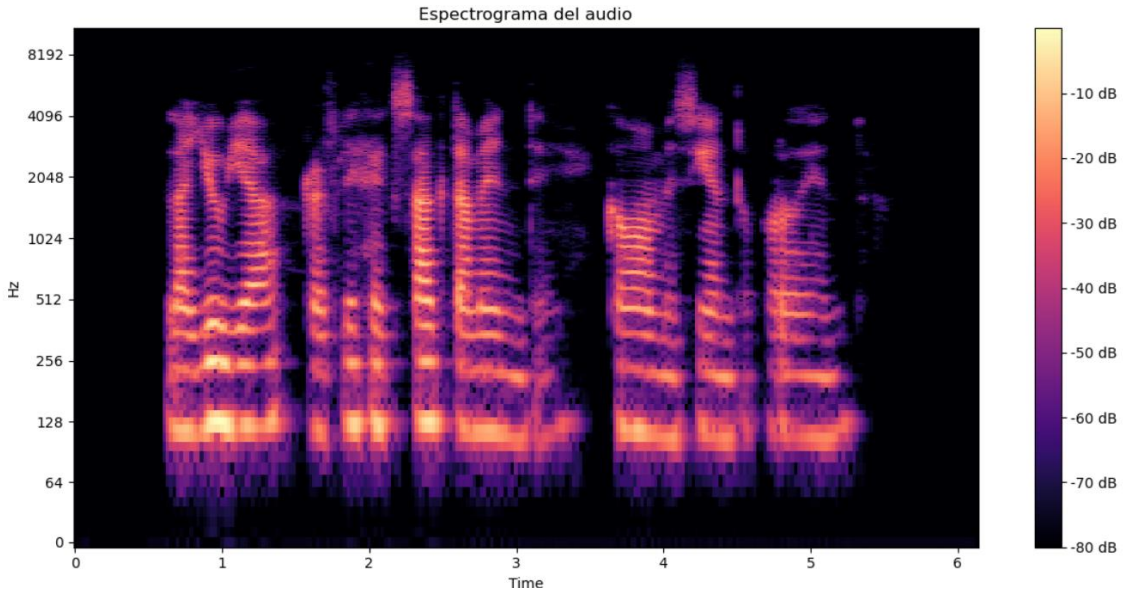


Ilustración 20. Espectograma de audio de tristeza. Elaboración propia

Luego de desarrollar todas las etapas y procesos previos del reconocimiento de la información sobre los audios, se presenta el modelo propuesto y configurado adecuadamente para el reconocimiento de emociones.

```
model = models.Sequential([
    layers.Input(shape=input_shape),

    # Primera capa convolucional - extrae patrones frecuenciales
    layers.Conv2D(32, (3, 3), padding='same', activation='relu'),
    layers.BatchNormalization(),
    layers.MaxPooling2D((2, 2), padding='same'),
    layers.Dropout(0.3),

    # Segunda capa convolucional - extrae patrones más complejos
    layers.Conv2D(64, (3, 3), padding='same', activation='relu'),
    layers.BatchNormalization(),
    layers.MaxPooling2D((2, 2), padding='same'),
```

```
    layers.Dropout(0.3),

    # Tercera capa convolucional
    layers.Conv2D(128, (3, 3), padding='same', activation='relu'),
    layers.BatchNormalization(),
    layers.MaxPooling2D((2, 2), padding='same'),
    layers.Dropout(0.4),

    # Aplanar para capas densas
    layers.Flatten(),

    # Capa densa con tamaño ajustado dinámicamente
    layers.Dense(256, activation='relu'),
    layers.Dropout(0.3),

    # Capa de salida
    layers.Dense(len(EMOTIONS), activation='softmax')
])

model.compile(
    optimizer=tf.keras.optimizers.Adam(learning_rate=0.0005),
    loss='sparse_categorical_crossentropy',
    metrics=['accuracy']
)
```

Arquitectura del modelo

Se diseñó una red neuronal convolucional profunda utilizando Keras con TensorFlow como backend. El modelo fue estructurado como una pila de capas Sequential, optimizada para el reconocimiento visual de patrones en espectrogramas.

Sección	Tipo de Capa	Descripción técnica	Función específica
Entrada	Input (shape=input_shape)	Entrada de dimensión (altura, ancho, canales) correspondiente a la imagen MFCC	Recepción de datos normalizados derivados del audio
1ª convolución	Conv2D(32, 3x3) + BatchNorm + MaxPooling + Dropout(0.3)	Extrae bordes, frecuencias simples; estabiliza y reduce dimensionalidad	Detecta patrones frecuenciales básicos como cambios tonales o de energía
2ª convolución	Conv2D(64, 3x3) + BatchNorm + MaxPooling + Dropout(0.3)	Extrae patrones compuestos a partir de filtros previos	Reconoce configuraciones espectro-temporales complejas
3ª convolución	Conv2D(128, 3x3) + BatchNorm + MaxPooling + Dropout(0.4)	Extrae estructuras profundas; alto nivel de abstracción	Permite discriminar emociones similares mediante señales más finas
Aplanamiento	Flatten()	Convierte el mapa de características 2D a vector 1D	Prepara datos para la clasificación final
Capa densa	Dense(256, relu) + Dropout(0.3)	Realiza combinaciones no lineales	Integra todas las características extraídas para decidir
Salida	Dense(len(EMOTIONS), softmax)	Produce una probabilidad por emoción	Clasifica la emoción más probable salida categórica

Tabla 3. Descripción de la arquitectura de la red. Elaboración propia

Compilación y parámetros de entrenamiento

Optimizador: Adam, adaptativo, eficiente para problemas no lineales.

Tasa de aprendizaje: 0.0005, elegida para garantizar convergencia estable.

Función de pérdida: sparse_categorical_crossentropy, ideal para clasificación multiclase con etiquetas enteras.

Métrica de evaluación: accuracy, apropiada dado que las clases son mutuamente exclusivas.

Justificación del diseño

El uso de capas convolucionales progresivas permite identificar patrones jerárquicos en el dominio espectral del habla.

La normalización por lotes (BatchNormalization) mejora la estabilidad del entrenamiento y reduce problemas de covariante shift.

Dropout actúa como regularizador para prevenir el sobreajuste, especialmente crítico cuando se trabaja con datasets moderados.

El modelo incluye un número creciente de filtros (32, 64 y 128), lo cual incrementa la capacidad de representación sin aumentar la profundidad excesiva.

La última capa densa utiliza softmax para asegurar una salida de probabilidad en las seis emociones.

Ventajas

Modelo adecuado para dispositivos de cómputo moderados (Colab GPU).

Bajo riesgo de sobreajuste debido a dropout + batch normalization.

4.4. Resultados

Descrito el proceso a realizar en la etapa de desarrollo, se procedió a realizar la programación correspondiente y a realizar el entrenamiento del modelo, para obtener los siguientes resultados.

Ejecución de la red:

```
- Estadísticas para Alegria:
  - Archivos originales: 779
  - Archivos válidos (duración correcta): 246
Porcentaje válidos: 31.6%
  - Necesarios 4 archivos adicionales por data augmentation
  - Estadísticas para Enojo:
  - Archivos originales: 772
  - Archivos válidos (duración correcta): 406
Porcentaje válidos: 52.6%
No se necesitó data augmentation
  - Estadísticas para Miedo:
  - Archivos originales: 766
  - Archivos válidos (duración correcta): 367
Porcentaje válidos: 47.9%
No se necesitó data augmentation
  - Estadísticas para Neutral:
  - Archivos originales: 778
  - Archivos válidos (duración correcta): 332
Porcentaje válidos: 42.7%
No se necesitó data augmentation
  - Estadísticas para Sorpresa:
  - Archivos originales: 772
  - Archivos válidos (duración correcta): 396
Porcentaje válidos: 51.3%
No se necesitó data augmentation
  - Estadísticas para Tristeza:
  - Archivos originales: 781
  - Archivos válidos (duración correcta): 234
Porcentaje válidos: 30.0%
  - Necesarios 16 archivos adicionales por data augmentation
Resumen estadístico de audios por emoción:
Emoción      | Originales | Válidos | Final
-----------------------------------------------
Alegria      |    779     |   246   |  250
Enojo        |    772     |   406   |  250
Miedo        |    766     |   367   |  250
Neutral      |    778     |   332   |  250
Sorpresa     |    772     |   396   |  250
Tristeza     |    781     |   234   |  250
Porcentaje de archivos válidos utilizados:
Alegria     : 246/246 (100.0% usado) + 4 aumentados
Enojo       : 250/406 (61.6% usado) + 0 aumentados
Miedo       : 250/367 (68.1% usado) + 0 aumentados
Neutral     : 250/332 (75.3% usado) + 0 aumentados
Sorpresa    : 250/396 (63.1% usado) + 0 aumentados
Tristeza    : 234/234 (100.0% usado) + 16 aumentados
```

```
Forma de los datos de entrada: (1500, 13, 137, 3)
Distribución de clases: {'Alegria': 250, 'Enojo': 250, 'Miedo': 250, 'Neu-
tral': 250, 'Sorpresa': 250, 'Tristeza': 250}
Todos los espectrogramas tienen la misma forma: (13, 137, 3)
Model: "sequential_3"
```

Layer (type)	Output Shape	Param #
conv2d_9 (Conv2D)	(None, 13, 137, 32)	896
batch_normalization_9 (BatchNormalization)	(None, 13, 137, 32)	128
max_pooling2d_9 (MaxPooling2D)	(None, 7, 69, 32)	0
dropout_12 (Dropout)	(None, 7, 69, 32)	0
conv2d_10 (Conv2D)	(None, 7, 69, 64)	18,496
batch_normalization_10 (BatchNormalization)	(None, 7, 69, 64)	256
max_pooling2d_10 (MaxPooling2D)	(None, 4, 35, 64)	0
dropout_13 (Dropout)	(None, 4, 35, 64)	0
conv2d_11 (Conv2D)	(None, 4, 35, 128)	73,856
batch_normalization_11 (BatchNormalization)	(None, 4, 35, 128)	512
max_pooling2d_11 (MaxPooling2D)	(None, 2, 18, 128)	0
dropout_14 (Dropout)	(None, 2, 18, 128)	0
flatten_3 (Flatten)	(None, 4608)	0
dense_6 (Dense)	(None, 256)	1,179,904
dropout_15 (Dropout)	(None, 256)	0
dense_7 (Dense)	(None, 6)	1,542

Tabla 4. Resumen de la red convolucional. Elaboración propia

Total params: 1,275,590 (4.87 MB)

Trainable params: 1,275,142 (4.86 MB)

Non-trainable params: 448 (1.75 KB)

Entrenamiento de la red (50 Epochs)

```
Epoch 1/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 9s 60ms/step - accuracy: 0.1887 - loss: 3.6567 -
val_accuracy: 0.1700 - val_loss: 1.7790
Epoch 2/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 55ms/step - accuracy: 0.2426 - loss: 1.8738 -
val_accuracy: 0.1667 - val_loss: 2.2264
Epoch 3/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 47ms/step - accuracy: 0.2667 - loss: 1.8336 -
val_accuracy: 0.1667 - val_loss: 2.9498
Epoch 4/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 48ms/step - accuracy: 0.2871 - loss: 1.7624 -
val_accuracy: 0.1667 - val_loss: 3.1972
Epoch 5/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.2962 - loss: 1.7490 -
val_accuracy: 0.1633 - val_loss: 3.2059
Epoch 6/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 53ms/step - accuracy: 0.3265 - loss: 1.6934 -
val_accuracy: 0.1667 - val_loss: 2.9144
Epoch 7/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.3618 - loss: 1.6249 -
val_accuracy: 0.1700 - val_loss: 2.7808
Epoch 8/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 56ms/step - accuracy: 0.3604 - loss: 1.6448 -
val_accuracy: 0.1967 - val_loss: 2.2021
Epoch 9/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 51ms/step - accuracy: 0.3849 - loss: 1.6023 -
val_accuracy: 0.1900 - val_loss: 2.2008
Epoch 10/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 48ms/step - accuracy: 0.3848 - loss: 1.5623 -
val_accuracy: 0.2467 - val_loss: 1.9099
Epoch 11/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.4334 - loss: 1.4793 -
val_accuracy: 0.3767 - val_loss: 1.6279
Epoch 12/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.4653 - loss: 1.4375 -
val_accuracy: 0.4300 - val_loss: 1.5738
Epoch 13/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 59ms/step - accuracy: 0.4994 - loss: 1.3703 -
val_accuracy: 0.4533 - val_loss: 1.4309
Epoch 14/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 50ms/step - accuracy: 0.5188 - loss: 1.3284 -
val_accuracy: 0.4600 - val_loss: 1.4050
Epoch 15/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 55ms/step - accuracy: 0.4824 - loss: 1.3692 -
val_accuracy: 0.5200 - val_loss: 1.2909
Epoch 16/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 52ms/step - accuracy: 0.5319 - loss: 1.2777 -
val_accuracy: 0.5400 - val_loss: 1.2623
Epoch 17/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 50ms/step - accuracy: 0.5558 - loss: 1.1997 -
val_accuracy: 0.5400 - val_loss: 1.2560
Epoch 18/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.5940 - loss: 1.1321 -
val_accuracy: 0.6000 - val_loss: 1.1042
Epoch 19/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 51ms/step - accuracy: 0.6178 - loss: 1.1002 -
val_accuracy: 0.6033 - val_loss: 1.1207
Epoch 20/50
```

```
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 52ms/step - accuracy: 0.6168 - loss: 1.1074 -
val_accuracy: 0.6300 - val_loss: 1.0713
Epoch 21/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 50ms/step - accuracy: 0.6471 - loss: 0.9721 -
val_accuracy: 0.6633 - val_loss: 1.0282
Epoch 22/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 58ms/step - accuracy: 0.6760 - loss: 0.9217 -
val_accuracy: 0.6533 - val_loss: 1.0316
Epoch 23/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 52ms/step - accuracy: 0.6932 - loss: 0.8916 -
val_accuracy: 0.6500 - val_loss: 1.0243
Epoch 24/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 48ms/step - accuracy: 0.6718 - loss: 0.8973 -
val_accuracy: 0.6900 - val_loss: 0.9792
Epoch 25/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 48ms/step - accuracy: 0.7049 - loss: 0.8132 -
val_accuracy: 0.6767 - val_loss: 1.0362
Epoch 26/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 48ms/step - accuracy: 0.7144 - loss: 0.7563 -
val_accuracy: 0.6833 - val_loss: 0.9877
Epoch 27/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.7492 - loss: 0.7453 -
val_accuracy: 0.7000 - val_loss: 0.8929
Epoch 28/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 52ms/step - accuracy: 0.7778 - loss: 0.6636 -
val_accuracy: 0.7367 - val_loss: 0.9063
Epoch 29/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 3s 73ms/step - accuracy: 0.7735 - loss: 0.6316 -
val_accuracy: 0.7033 - val_loss: 0.9600
Epoch 30/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 3s 70ms/step - accuracy: 0.7785 - loss: 0.6182 -
val_accuracy: 0.6933 - val_loss: 0.9663
Epoch 31/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 62ms/step - accuracy: 0.8422 - loss: 0.4404 -
val_accuracy: 0.7167 - val_loss: 0.9387
Epoch 32/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 54ms/step - accuracy: 0.7938 - loss: 0.5228 -
val_accuracy: 0.7233 - val_loss: 0.9413
Epoch 33/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 64ms/step - accuracy: 0.8328 - loss: 0.4632 -
val_accuracy: 0.7233 - val_loss: 0.9340
Epoch 34/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 53ms/step - accuracy: 0.8550 - loss: 0.3920 -
val_accuracy: 0.7100 - val_loss: 0.9186
Epoch 35/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 61ms/step - accuracy: 0.8523 - loss: 0.4422 -
val_accuracy: 0.7300 - val_loss: 0.9217
Epoch 36/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.8943 - loss: 0.3291 -
val_accuracy: 0.7433 - val_loss: 0.9132
Epoch 37/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 50ms/step - accuracy: 0.8798 - loss: 0.3429 -
val_accuracy: 0.7267 - val_loss: 0.9565
Epoch 38/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 61ms/step - accuracy: 0.9002 - loss: 0.3186 -
val_accuracy: 0.7367 - val_loss: 0.9622
Epoch 39/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 50ms/step - accuracy: 0.8970 - loss: 0.2965 -
val_accuracy: 0.7400 - val_loss: 0.9696
Epoch 40/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.9175 - loss: 0.2400 -
val_accuracy: 0.7233 - val_loss: 1.0184
```

```
Epoch 41/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 51ms/step - accuracy: 0.9162 - loss: 0.2527 -
val_accuracy: 0.7633 - val_loss: 0.9705
Epoch 42/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 54ms/step - accuracy: 0.9199 - loss: 0.2554 -
val_accuracy: 0.7600 - val_loss: 1.0286
Epoch 43/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 58ms/step - accuracy: 0.9114 - loss: 0.2372 -
val_accuracy: 0.7333 - val_loss: 0.9996
Epoch 44/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.9424 - loss: 0.1832 -
val_accuracy: 0.7667 - val_loss: 1.0103
Epoch 45/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.9344 - loss: 0.1924 -
val_accuracy: 0.7533 - val_loss: 1.0507
Epoch 46/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 49ms/step - accuracy: 0.9325 - loss: 0.1901 -
val_accuracy: 0.7767 - val_loss: 0.9723
Epoch 47/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 54ms/step - accuracy: 0.9418 - loss: 0.1787 -
val_accuracy: 0.7533 - val_loss: 1.0911
Epoch 48/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 54ms/step - accuracy: 0.9432 - loss: 0.1657 -
val_accuracy: 0.7667 - val_loss: 0.9967
Epoch 49/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 50ms/step - accuracy: 0.9592 - loss: 0.1321 -
val_accuracy: 0.7500 - val_loss: 1.0275
Epoch 50/50
38/38 ━━━━━━━━━━━━━━━━━━━━ 2s 50ms/step - accuracy: 0.9627 - loss: 0.1269 -
val_accuracy: 0.7633 - val_loss: 0.9515
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 21ms/step
```

La evaluación del modelo se realizó sobre un conjunto de prueba compuesto por 300 muestras balanceadas, correspondientes a seis emociones: Alegría, Enojo, Miedo, Neutral, Sorpresa y Tristeza. El rendimiento del modelo fue medido mediante las métricas estándar de clasificación: precisión, recall y F1-score, además de la exactitud global (accuracy). Estas métricas indican un balance razonable entre las clases, sin una dominancia marcada por alguna categoría, ya que el conjunto de datos estaba balanceado por diseño.

```
Reporte de clasificación:
              precision    recall  f1-score   support

     Alegria       0.84      0.76      0.80        50
       Enojo       0.86      0.84      0.85        50
       Miedo       0.67      0.60      0.63        50
     Neutral       0.85      0.82      0.84        50
    Sorpresa       0.71      0.74      0.73        50
    Tristeza       0.67      0.82      0.74        50

    accuracy                           0.76       300
   macro avg       0.77      0.76      0.76       300
weighted avg       0.77      0.76      0.76       300
```

Tabla 5. Resultados del entrenamiento. Elaboración propia

Mejor desempeño: Se observó en las clases Neutral, Enojo y Alegría, con F1-scores promedio sobre 0.80, lo cual indica que el modelo logra identificar estas emociones de forma efectiva y consistente.

Desempeño intermedio: La emoción Tristeza obtuvo un desempeño también relativamente sólido (0.74 F1), lo que sugiere que esta emoción presenta patrones acústicos bien diferenciados.

Desafíos en la clasificación:

Miedo presentó el menor F1-score (0.63), lo que indica dificultades en la identificación de esta emoción, posiblemente debido a su similitud acústica con otras emociones como Sorpresa o Tristeza.

Miedo mostró un patrón interesante: recall cercano al alto (0.60). Esto sugiere que el modelo tiende a clasificar otras emociones como Sorpresa con frecuencia, lo cual podría deberse a rasgos acústicos compartidos (e.g. tono elevado, energía súbita).

A continuación, se detallan los resultados obtenidos.

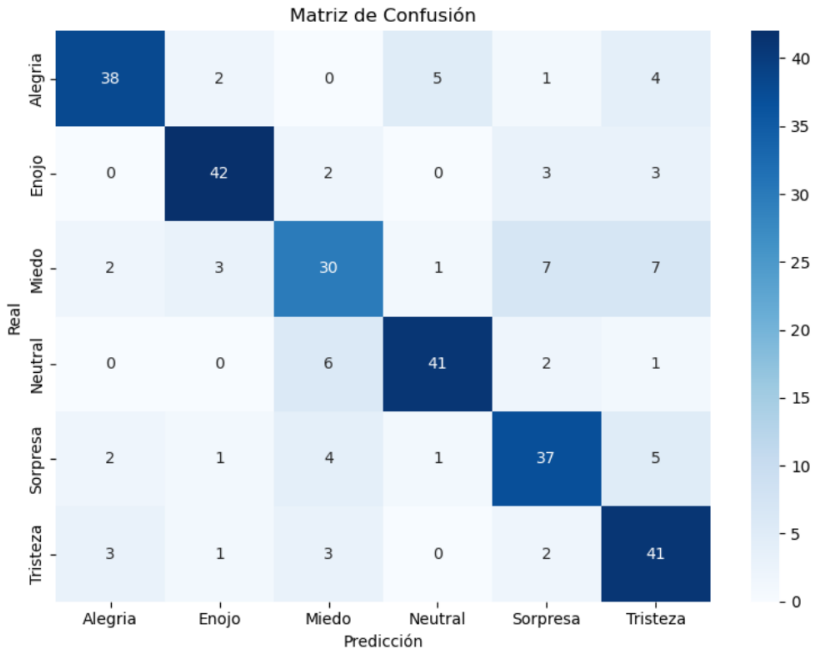


Ilustración 21. Matriz de confusión. Elaboración propia

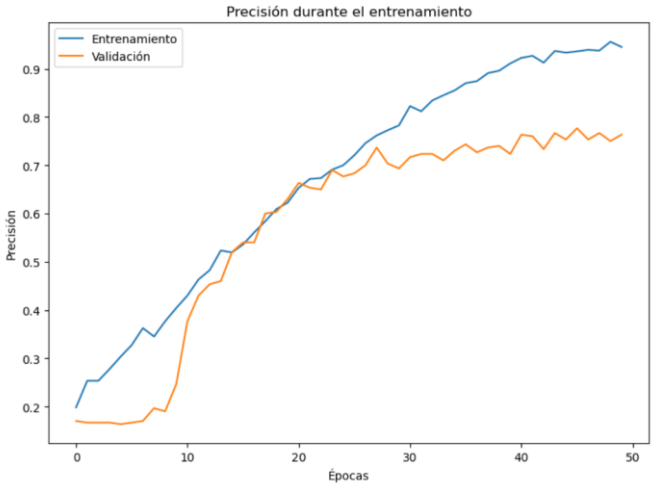


Ilustración 22. Precisión durante entrenamiento. Elaboración propia

Análisis técnico de las curvas.

1. Fase inicial (épocas 0–10):

   La precisión en entrenamiento comienza en aproximadamente 20% y aumenta progresivamente.

   La curva de validación se mantiene casi plana hasta la época 7, lo cual es común cuando el modelo aún no ha aprendido patrones generalizables.

   A partir de la época 8, ambas curvas se incrementan de forma coordinada.

2. Fase media (épocas 10–25):

   Se observa una convergencia temporal alrededor de la época 20, donde ambas curvas alcanzan valores similares.

   Esto sugiere que el modelo ha aprendido a representar correctamente las características relevantes sin sobreajustarse aún.

3. Fase final (épocas 25–50):

   La precisión de entrenamiento continúa en ascenso hasta superar el 90%.

   La precisión de validación se estabiliza cerca del 75%, lo que indica una ligera brecha de generalización, aunque sin evidencia de sobreajuste.

   No se observan oscilaciones violentas ni descensos abruptos, lo cual indica una entrenabilidad estable.

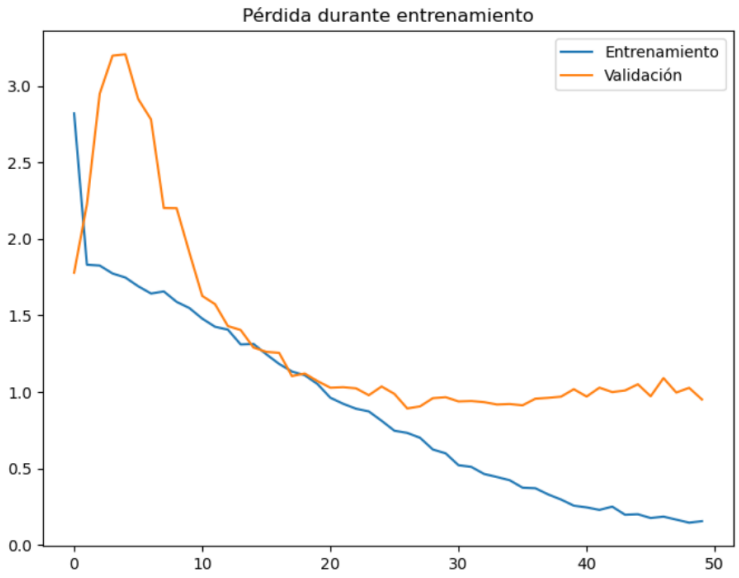


Ilustración 23. Pérdida durante el entrenamiento. Elaboración propia

Análisis técnico de las curvas.

1. Épocas 0–10:

   La pérdida de validación muestra una fluctuación alta, característica de modelos que aún no han logrado aprender generalización.

   Posible sensibilidad a datos iniciales o bajo número de muestras por clase en validación.

2. Épocas 10–20:

   La pérdida en validación se reduce abruptamente, reflejando la mejora de la precisión.

   Ambas curvas convergen brevemente alrededor de la época 20.

3. Épocas 20–50:

   La pérdida en entrenamiento sigue bajando, lo que muestra un ajuste continuo del modelo.

   La pérdida en validación se estabiliza y fluctúa levemente, lo cual indica el posible comienzo del sobreajuste: el modelo sigue mejorando en entrenamiento.

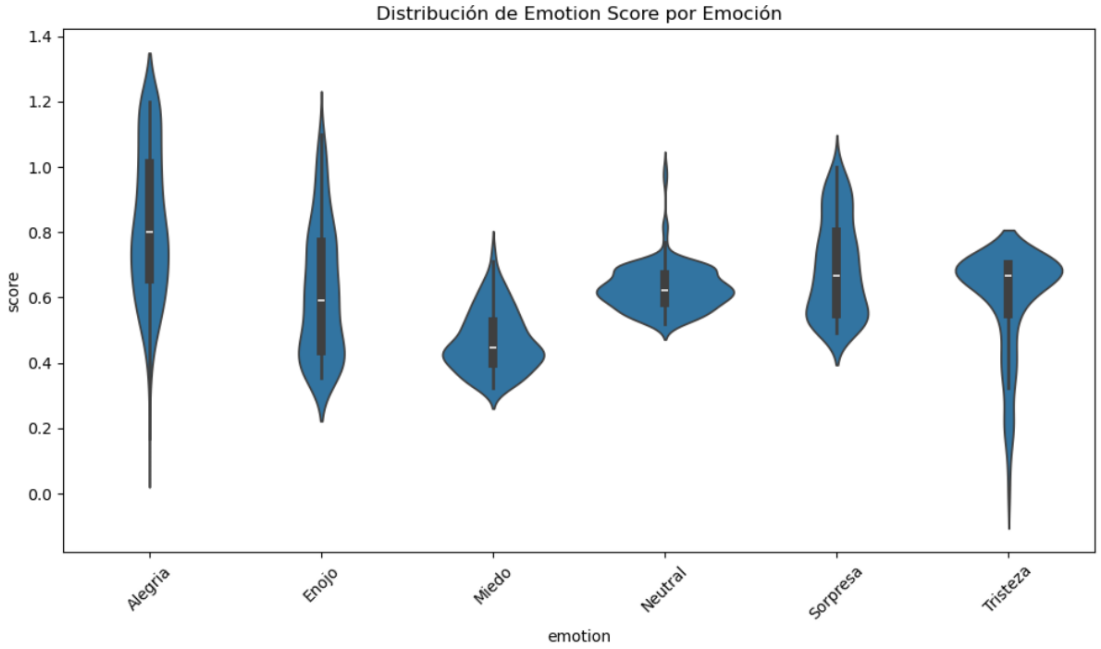


Ilustración 24. Datos agrupados por emoción. Elaboración propia

Emoción	Media / Mediana	Dispersión	Observaciones clave
Alegría	Alta (≥ 0.8)	Amplia	Presenta la mayor dispersión; hay valores que llegan hasta 1.3. La mediana es alta, lo que sugiere que los audios clasificados como alegría suelen tener altos scores emocionales.
Enojo	Moderada (~0.6)	Amplia	Score centrado, pero con dispersión hacia valores bajos y altos. Muestra cierta variabilidad en cómo se expresa acústicamente el enojo.
Miedo	Baja (~0.45)	Estrecha	Menor dispersión. Los valores están concentrados, lo que sugiere una representación acústica más homogénea para el miedo.
Neutral	Moderada (~0.6)	Controlada	Score bastante uniforme. La forma del violín es simétrica, lo que sugiere estabilidad en la representación de esta emoción.
Sorpresa	Moderada-alta (~0.7)	Amplia	Amplia distribución, similar a alegría, pero más centrada. Puede confundirse con emociones intensas como alegría o miedo.
Tristeza	Moderada (~0.6)	Alta	Amplia dispersión, con algunos valores cercanos a cero. Puede representar diferencias culturales o variabilidad vocal individual en la expresión de tristeza.

Tabla 6. Interpretación detallada por emoción. Elaboración propia

Evaluación del dataset con otros modelos.

A continuación, se muestran los resultados de las comparaciones realizadas con otros modelos que se utilizan para realizar análisis de emociones con el dataset que ya ha comprimido los audios.

Resultados Random Forest

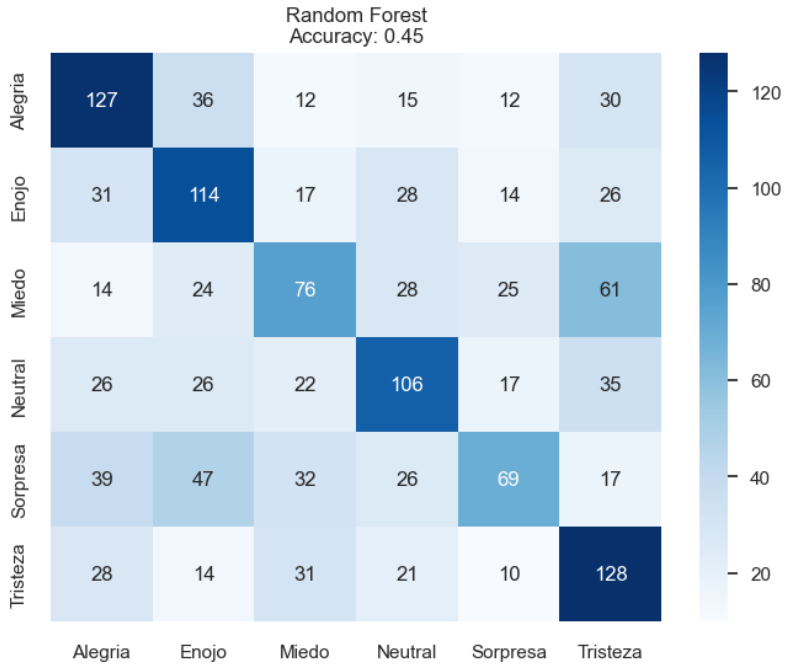


Ilustración 25. Matriz de confusión de Random Forest. Elaboración propia

Reporte Random Forest:

	precision	recall	f1-score	support
Alegria	0.48	0.55	0.51	232
Enojo	0.44	0.50	0.46	230
Miedo	0.40	0.33	0.36	228
Neutral	0.47	0.46	0.46	232
Sorpresa	0.47	0.30	0.37	230
Tristeza	0.43	0.55	0.48	232
accuracy			0.45	1384
macro avg	0.45	0.45	0.44	1384
weighted avg	0.45	0.45	0.44	1384

Tabla 7. Resultados sobre Random Forest. Elaboración propia

Resultados SVM (Support Vector Machine)

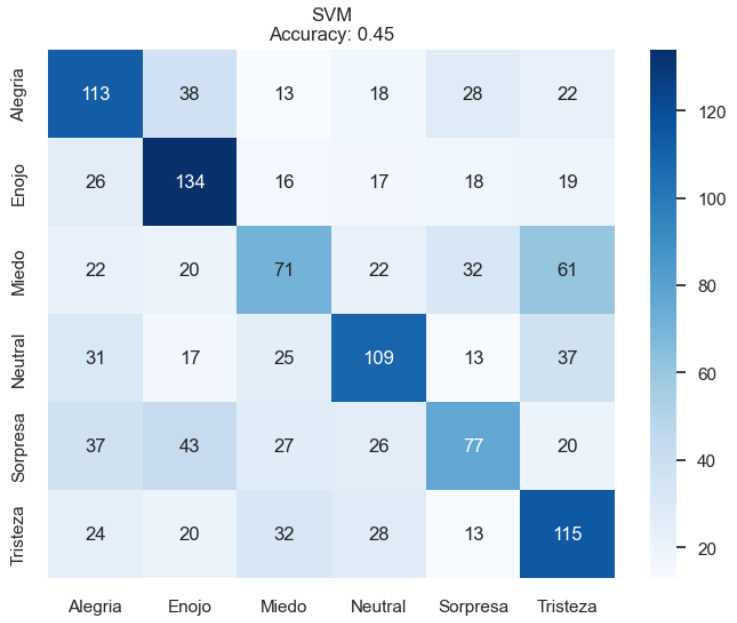


Ilustración 26. Resultados con SVM. Elaboración propia

Reporte SVM:

	precision	recall	f1-score	support
Alegria	0.45	0.49	0.47	232
Enojo	0.49	0.58	0.53	230
Miedo	0.39	0.31	0.34	228
Neutral	0.50	0.47	0.48	232
Sorpresa	0.43	0.33	0.37	230
Tristeza	0.42	0.50	0.45	232
accuracy			0.45	1384
macro avg	0.44	0.45	0.44	1384
weighted avg	0.44	0.45	0.44	1384

Tabla 8. Resultados con SVM. Elaboración propia

Resultados XGBoost

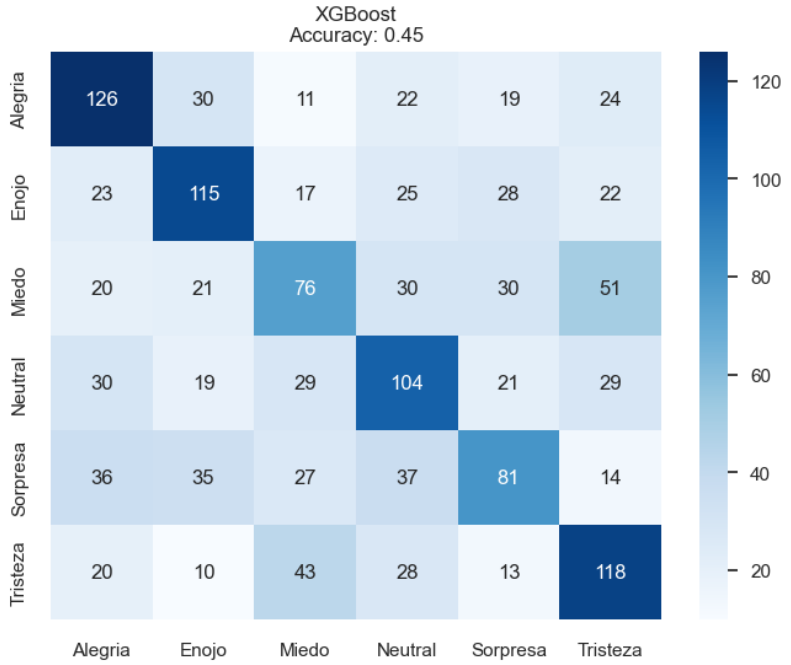


Ilustración 27. Resultados con XGBoost. Elaboración propia

Reporte XGBoost:

	precision	recall	f1-score	support
Alegria	0.49	0.54	0.52	232
Enojo	0.50	0.50	0.50	230
Miedo	0.37	0.33	0.35	228
Neutral	0.42	0.45	0.44	232
Sorpresa	0.42	0.35	0.38	230
Tristeza	0.46	0.51	0.48	232
accuracy			0.45	1384
macro avg	0.45	0.45	0.45	1384
weighted avg	0.45	0.45	0.45	1384

Tabla 9. Resultados con XGBoost. Elaboración propia

Resultados Multilayer Perceptron o Perceptrón Multicapa

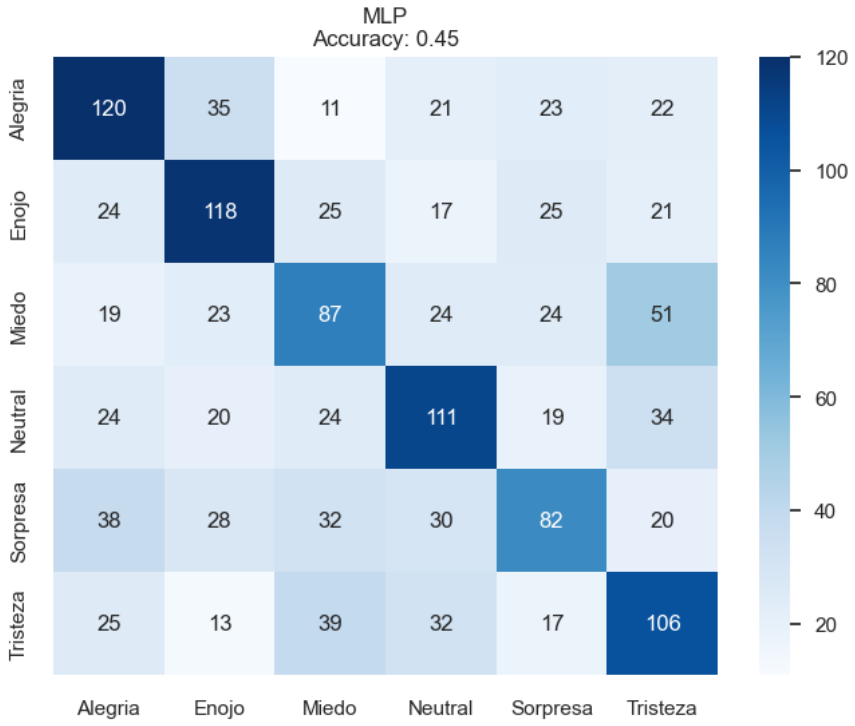


Ilustración 28. Resultados con MLP. Elaboración propia

Reporte MLP:

	precision	recall	f1-score	support
Alegria	0.48	0.52	0.50	232
Enojo	0.50	0.51	0.51	230
Miedo	0.40	0.38	0.39	228
Neutral	0.47	0.48	0.48	232
Sorpresa	0.43	0.36	0.39	230
Tristeza	0.42	0.46	0.44	232
accuracy			0.45	1384
macro avg	0.45	0.45	0.45	1384
weighted avg	0.45	0.45	0.45	1384

Tabla 10. Resultados con MLP. Elaboración propia

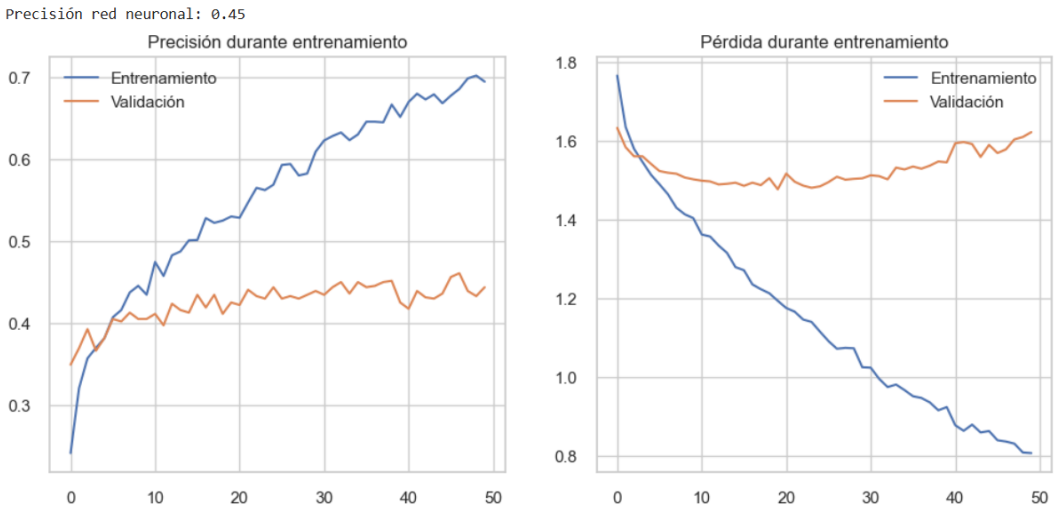


Ilustración 29. mejor modelo: MLP. Elaboración propia

Resultados YAMNet + Random Forest

El resultado del siguiente modelo esta dado por un modelo estándar para el reconocimiento de diversos tipos de audios, entre ellos las emociones, en donde, compara la intensidad, entre otros factores, para determinar el grado de similitud con las emociones registradas.

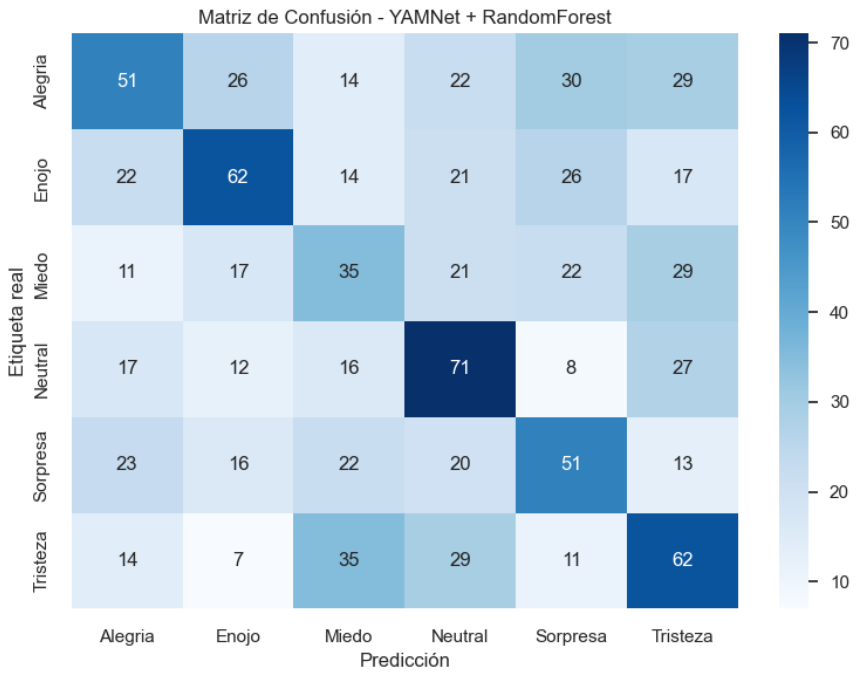


Ilustración 30. Modelo YamNet + Random Forest. Elaboración propia

--- Reporte de Clasificación ---

	precision	recall	f1-score	support
Alegria	0.37	0.30	0.33	172
Enojo	0.44	0.38	0.41	162
Miedo	0.26	0.26	0.26	135
Neutral	0.39	0.47	0.42	151
Sorpresa	0.34	0.35	0.35	145
Tristeza	0.35	0.39	0.37	158
accuracy			0.36	923
macro avg	0.36	0.36	0.36	923
weighted avg	0.36	0.36	0.36	923

Tabla 11. Resultados modelo YamNet + Random Forest. Elaboración propia

Resultados MFCC + STFT

Short-Time Fourier Transform (Transformada de Fourier de Tiempo Corto)

Del mismo modo, se realizó una comparación con el modelo completo de MFCC + STFT y Random Forest nuevamente, obteniéndose los siguientes resultados.

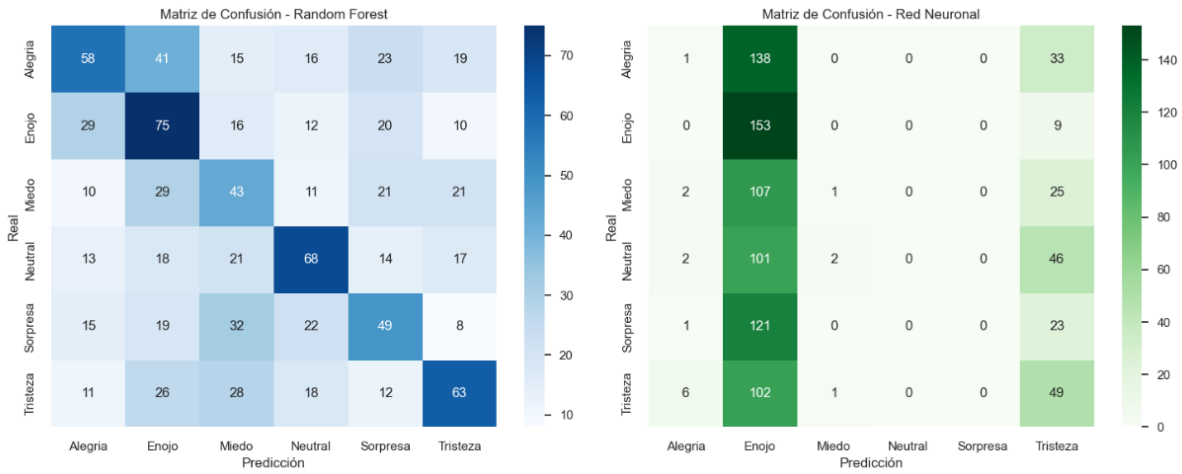


Tabla 12. Resultados modelo MFCC + STFT y Random Forest. Elaboración propia

Red Neuronal Convolucional:

	precision	recall	f1-score	support
Alegria	0.08	0.01	0.01	172
Enojo	0.21	0.94	0.35	162
Miedo	0.25	0.01	0.01	135
Neutral	0.00	0.00	0.00	151
Sorpresa	0.00	0.00	0.00	145
Tristeza	0.26	0.31	0.29	158
accuracy			0.22	923
macro avg	0.14	0.21	0.11	923
weighted avg	0.13	0.22	0.11	923

Ilustración 31. Resultados MFCC + STFT. Elaboración propia

Resultado modelo con sub bandas de audio

Modelo a comparar fue utilizando sub bandas dentro del manejo de audios, se muestran los detalles en los siguientes resultados.

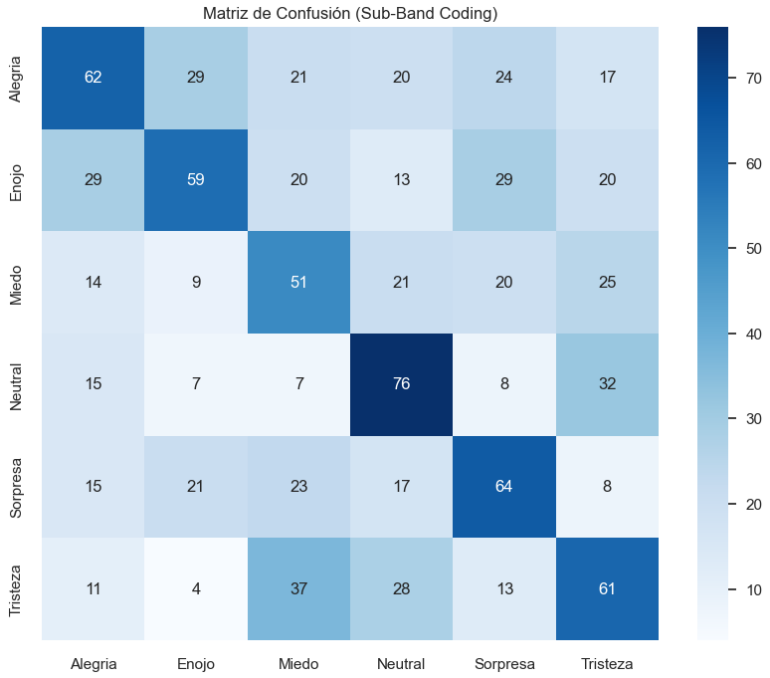


Ilustración 32. Resultados sub bandas de audios. Elaboración propia

-- Reporte de Clasificación (con SBC) ---

	precision	recall	f1-score	support
Alegria	0.42	0.36	0.39	173
Enojo	0.46	0.35	0.39	170
Miedo	0.32	0.36	0.34	140
Neutral	0.43	0.52	0.47	145
Sorpresa	0.41	0.43	0.42	148
Tristeza	0.37	0.40	0.38	154
accuracy			0.40	930
macro avg	0.40	0.40	0.40	930
weighted avg	0.41	0.40	0.40	930

Tabla 13. Resultados sub bandas de audios. Elaboración propia

Resultados de CNN + LSTM

Por último, se realiza una prueba con el modelo indicado, esperando pueda ser una mejor solución, sin embargo, el proceso se comienza a caerá desde el inicio, no pudiendo llegar a los resultados ideales. Se muestran los resultados obtenidos.

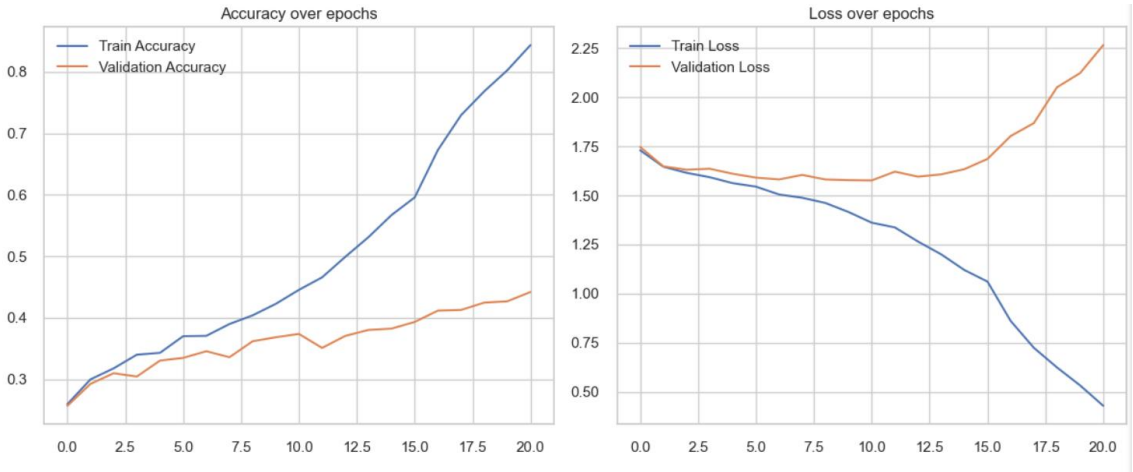


Ilustración 33. Resultados modelo CNN + LSTM. Elaboración propia

Classification Report:

	precision	recall	f1-score	support
Alegria	0.49	0.34	0.40	155
Enojo	0.35	0.44	0.39	153
Miedo	0.33	0.28	0.30	152
Neutral	0.32	0.40	0.35	155
Sorpresa	0.52	0.29	0.37	153
Tristeza	0.35	0.48	0.41	155
accuracy			0.37	923
macro avg	0.39	0.37	0.37	923
weighted avg	0.39	0.37	0.37	923

Tabla 14. Resultados modelo CNN + LSTM. Elaboración propia

Con lo cual, se puede apreciar después de las evaluaciones realizadas, que el modelo propuesto tiene mejores resultados frente a otros modelos utilizando el mismo conjunto de datos con los audios que ya cuentan con compresión.

5. Conclusión y trabajos futuros

Conclusiones

De acuerdo con lo desarrollado en la presente investigación, se concluye lo siguiente:

- Se logró recolectar un conjunto de datos de voz etiquetados por emociones representativas del contexto académico universitario de una universidad de la ciudad de Lima-Perú, llegando a un total de 4,720 audios relacionados con las emociones de alegría, tristeza, neutral, enojo, sorpresa y miedo.

- Se logró implementar algoritmos de procesamiento de voz en Python para la extracción de características, principalmente con MFCC, compresión de audio, entre otros, descritos en el presente trabajo realizado.

- Se logró entrenar un modelo a partir de un conjunto de los audios recolectados para identificar emociones en los estudiantes ingresantes a una universidad pública en la ciudad de Lima-Perú, modelo que cuenta con características ideales para esta ciudad por su forma y tono de voz.

- Se logró evaluar el rendimiento del modelo desarrollado, mediante métricas de precisión, obteniéndose un 96% de entrenamiento, y un 76% de precisión, las diferencias encontradas se fundamentan en momentos en los cuales una emoción se relaciona a otra por el mismo grado de intensidad.

Trabajos futuros

En base al trabajo realizado se recomienda para los trabajos futuros lo siguiente:

Ampliar la base de datos de voces

Es recomendable aumentar la cantidad y diversidad de registros de voz, incluyendo:

- Participantes de diferentes edades, géneros y regiones del Perú.
- Contextos emocionales más variados (no solo académicos).
  Esto permitirá construir un modelo más robusto y generalizable a situaciones reales fuera del entorno universitario.

Incorporar técnicas avanzadas de augmentación de datos

Se sugiere aplicar métodos de data augmentation más sofisticados, como:

- Perturbaciones de pitch y tiempo controladas.
- Simulación de ruido ambiental real.
- Síntesis de voz emocional (con herramientas de TTS adaptativo).
  Esto ayudaría a mejorar la precisión sin necesidad de recolectar más datos reales.

Comparar arquitecturas de modelos

Se recomienda comparar el desempeño del modelo actual con otras arquitecturas como:

- Transformers acústicos (e.g., wav2vec2, AST).
- Modelos preentrenados ajustados al español latino.
  Esto permitirá identificar arquitecturas más adecuadas para reconocimiento de emociones en contextos peruanos.

Explorar nuevas características acústicas

Además de los MFCC, sería útil incorporar:

- Prosodia (duración de sílabas, pausas, entonación).
- Características subbandales y espectro-temporales.
- Análisis no lineales como jitter, shimmer, y formantes vocales.

Desarrollar una aplicación práctica en tiempo real

Se recomienda como siguiente etapa el diseño de una **interfaz interactiva**, por ejemplo:

- Un módulo de detección de emociones en clases virtuales.
- Integración en sistemas de orientación psicológica o tutorías.
- Herramientas de retroalimentación emocional para docentes.

Estudiar correlación con variables contextuales

En futuros estudios puede ser útil analizar cómo las emociones detectadas se relacionan con:

- Nivel de estrés académico.
- Desempeño en cursos.
- Participación en clases virtuales o presenciales.

6. Referencias

Ahmed Al Kuwaiti, Khalid Nazer, Abdullah H. Alreedy, Shaher D AlShehri, Afnan Almuhanna, Arun Vijay Subbarayalu, Dhoha Al Muhanna, et al. (2023). A Review of the Role of Artificial Intelligence in Healthcare. Volume(13), 951-951. Journal of Personalized Medicine. https://doi.org/10.3390/jpm13060951

Atassi, Hicham (2014) Emotion Recognition from Acted and Spontaneous Speech. doi: https://core.ac.uk/download/30311129.pdf

Azzopardi, Leif, Halvey, Martin, Macdonald, Craig, Ounis, et al. (2017) Report on the Information Retrieval Festival (IRFest2017). doi: https://core.ac.uk/download/96881654.pdf

Ball, Leslie D., Bradley, David A., Brownsell, Simon, Szymkowiak, et al. (2011) Linking recorded data with emotive and adaptive computing in an eHealth environment. doi: https://core.ac.uk/download/228176557.pdf

Broek, Egon L. van den (2005) Empathic Agent Technology (EAT). doi: https://core.ac.uk/download/pdf/11482531.pdf

Cross, Emily S., Hekele, Felix, Hortensius, Ruud (2018) The perception of emotion in artificial agents. doi: https://core.ac.uk/download/157582956.pdf

Dethlefs, Brent A, Lee, Katherine L, Li, Shengwen Calvin, Loudon, et al. (2019) Breathing Signature as Vitality Score Index Created by Exercises of Qigong: Implications of Artificial Intelligence Tools Used in Traditional Chinese Medicine. doi: https://core.ac.uk/download/323075335.pdf

Firuz Kamalov, David Santandreu Calonge, Ikhlaas Gurrib (2023). New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. Volume(15), 12451-12451. Sustainability. https://doi.org/10.3390/su151612451

Gilleran, Brady (2019) Identificación de noticias falsas mediante el análisis de emociones. doi: https://core.ac.uk/download/215464206.pdf

Grawemeyer, Beate, Gutiérrez-Santos, Sergio, Holmes, Wayne, Mavrikis, et al. (2017) Aprendizaje afectivo: mejorar la participación y el aprendizaje con retroalimentación consciente del afecto. doi: https://core.ac.uk/download/82983953.pdf

Mohamed M. Abd ElMaksoud, A. H. Kandil, Sherif H. El-Gohary (2024) Staged Transfer Learning for Multi-Label Facial Emotion Recognition from Full Faces. 2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES). doi: h10.1109/NILES63360.2024.10753183

Pawan Budhwar, Soumyadeb Chowdhury, Geoffrey Wood, Herman Aguinis, Greg J. Bamber, Jose R. Beltran, Paul Boselie, et al. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. Volume(33), 606659. Human Resource Management Journal. https://doi.org/10.1111/1748-8583.12524

Qinglin Yang, Yetong Zhao, Huawei Huang, Zehui Xiong, Jiawen Kang, Zibin Zheng (2022). Fusing Blockchain and AI With Metaverse: A Survey. Volume(3), 122-136. IEEE Open Journal of the Computer Society. https://doi.org/10.1109/ojcs.2022.3188249

Reabal Najjar (2023). Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. Volume(13), 2760-2760. Diagnostics. https://doi.org/10.3390/diagnostics13172760

Sangmin Park, Young-Gab Kim (2022). A Metaverse: Taxonomy, Components, Applications, and open Challenges. Volume(10), 4209-4251. IEEE Access. https://doi.org/10.1109/access.2021.3140175

Sarah Bankins, Anna Carmella Ocampo, Mauricio Marrone, Simon Lloyd D. Restubog, Sang Eun Woo (2023). A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. Volume(45), 159-182. Journal of Organizational Behavior. https://doi.org/10.1002/job.2735

Shiva Maleki Varnosfaderani, Mohamad Forouzanfar (2024). The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century. Volume(11), 337-337. Bioengineering. https://doi.org/10.3390/bioengineering11040337

Sivek, Susan Currie (2018) Análisis ubicuo de emociones y cómo nos sentimos hoy. doi: https://core.ac.uk/download/212894256.pdf

Soelistio, Yustinus Eko, Wunarso, Novita Belinda (2017) Hacia el reconocimiento automático del habla y las emociones de Indonesia (I-SpEAR). doi: http://arxiv.org/abs/1709.10460

Urquhart, L., Miranda, D., Connon, I. L. C., & Laffer, A. (2023). Critically Envisioning Biometric Artificial Intelligence in Law Enforcement. University of Edinburgh.

Yogesh K. Dwivedi, Laurie Hughes, Arpan Kumar Kar, Abdullah M. Baabdullah, Purva Grover, Roba Abbas, Daniela Andreini, et al. (2021) Climate change and COP26: Are digital technologies and information management part of the problem or the solution? An editorial reflection and call to action. Volume(63), 102456-102456. International Journal of Information Management. doi: https://doi.org/10.1016/j.ijinfomgt.2021.102456.

Yogesh K. Dwivedi, Laurie Hughes, Abdullah M. Baabdullah, Samuel Ribeiro-Navarrete, Mihalis Giannakis, Mutaz M. Al-Debei, Denis Dennehy, et al. (2022) Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Volume(66), 102542-102542. International Journal of Information Management. doi: https://doi.org/10.1016/j.ijinfomgt.2022.102542.

Yogesh K. Dwivedi, Nir Kshetri, Laurie Hughes, Emma Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah, et al. (2023) Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Volume(71), 102642-102642. International Journal of Information Management. doi: https://doi.org/10.1016/j.ijinfomgt.2023.102642

Apéndice I

El apéndice es un adjunto al documento académico de autoría propia. No es un documento independiente, pues no se entendería si no es en relación con el resto del trabajo. Contiene información que complementa o aclara la tesis y que se considera que es demasiado larga o detallada para incluirse en el texto principal. Dicha información podría incluir gráficos o tablas, listas de datos sin procesar, etc.

Anexos I

Los anexos también contienen información adicional que se considera relevante para justificar las conclusiones del trabajo, pero, por lo general, el autor de contenido del anexo es distinto al autor del trabajo. Suele ser un documento independiente del trabajo. Pueden ser tablas de datos, imágenes, etc. Es necesario incluir las referencias de los documentos de donde procedan.