
**UNIVERSIDAD TÉCNICA DEL NORTE**

**FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS**

**CARRERA DE INGENIERÍA EN TELECOMUNICACIONES**

# “SISTEMA DE RECONOCIMIENTO DE EMOCIONES A TRAVÉS DE LA VOZ, MEDIANTE TÉCNICAS DE APRENDIZAJE PROFUNDO”

**TRABAJO DE GRADO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERÍA EN TELECOMUNICACIONES**

**AUTOR:** GUERRÓN PANTOJA CARLOS FERNANDO

**DIRECTOR:** ING. MAYA OLALLA EDGAR ALBERTO, MSC.

**ASESOR:** ING. ANA CRISTINA UMAQUINGA CRIOLLO, MSC.

**Ibarra-Ecuador**

**2023**

# UNIVERSIDAD TECNICA DEL NORTE
# BIBLIOTECA UNIVERSITARIA

## AUTORIZACIÓN DE USO Y PUBLICACIÓN
## A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE

**IDENTIFICACIÓN DE LA OBRA.**

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición la siguiente información:

| DATOS DEL CONTACTO | | | |
|---|---|---|---|
| **CÉDULA DE IDENTIDAD** | 0402074942 | | |
| **APELLIDOS Y NOMBRES** | Guerrón Pantoja Carlos Fernando | | |
| **DIRECCIÓN** | Ibarra | | |
| **E-MAIL** | cfguerronp@utn.edu.ec | | |
| **TELÉFONO FIJO** | 062250972 | **TELÉFONO MÓVIL** | 0984928504 |

| DATOS DE LA OBRA | |
|---|---|
| **TÍTULO** | "Sistema de reconocimiento de emociones a través de la voz, mediante técnicas de aprendizaje profundo" |
| **AUTOR** | Guerrón Pantoja Carlos Fernando |
| **FECHA** | 06/06/2023 |
| **PROGRAMA** | ■ **PREGRADO** ☐ **POSGRADO** |
| **TÍTULO** | Ingeniero en Telecomunicaciones |
| **DIRECTOR** | MSc. Maya Olalla Edgar Alberto |

**CONSTANCIAS**

El autor manifiesta que la obra objeto de la presente autorización es original y se la desarrolló, sin violar derechos de autor de terceros, por lo tanto, la obra es original y que es el titular de los derechos patrimoniales, por lo que asume la responsabilidad sobre el contenido de la misma y saldrá en defensa de la Universidad en caso de reclamación por parte de terceros.

Ibarra, a los 6 días del mes de junio de 2023

EL AUTOR

Guerrón Pantoja Carlos Fernando

CI: 0402074942

**UNIVERSIDAD TÉCNICA DEL NORTE**

**FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS**

**CERTIFICACIÓN**

MAGISTER MAYA OLALLA EDGAR ALBERTO, CON CÉDULA DE IDENTIDAD Nro. 1002702197, DIRECTOR DEL PRESENTE TRABAJO DE TITULACIÓN CERTIFICA:

Que, el presente trabajo de Titulación "SISTEMA DE RECONOCIMIENTO DE EMOCIONES A TRAVÉS DE LA VOZ, MEDIANTE TÉCNICAS DE APRENDIZAJE PROFUNDO" Ha sido desarrollado por el señor Guerrón Pantoja Carlos Fernando bajo mi supervisión.

Es todo en cuanto puedo certificar en honor de la verdad.

.....

Ing. Maya Olalla Edgar Alberto, MSc.

DIRECTOR

## DEDICATORIA

Dedico el presente proyecto de titulación a mi familia que siempre estuvo apoyándome para alcanzar este objetivo tan anhelado, en especial a mis padres, Carlos y Laura, ya que gracias a sus enseñanzas supieron guiarme en el transcurso de mi vida universitaria, siempre confiaron en mis capacidades y me alentaron para nunca rendirme.

*Carlos Fernando Guerrón Pantoja*

## AGRADECIMEINTO

En primer lugar, agradezco a Dios, ya que el me brindó sabiduría, conocimiento, entendimiento, dedicación, salud y bendecirme con la vida, también con una familia unida y amorosa. El me dio las fuerzas para seguir adelante y me cuido en cada viaje que realizaba desde mi ciudad natal "Tulcán" hacia la ciudad donde adquiría conocimientos y me preparaba para ser un profesional, sé que el siempre estará a mi lado cuidándome y bendiciéndome.

Agradezco el apoyo incondicional de mi familia, a mi padre que trabaja fuertemente para poder brindarme la educación, a mi madre que con su amor siempre me alentaba desde lejos y cuidaba de mi bienestar, a mi hermana Amanda que me abrió las puertas de su hogar y me brindo su hospitalidad para tener un lugar donde vivir en mi vida estudiantil, a mis hermanas Mireya, Gissela y Belén que siempre se preocupaban por mí y me brindaron el apoyo total a lo largo de mi carrera.

Agradezco a la Universidad Técnica del Norte por permitirme prepárame como profesional, a mi director MSc. Edgar Maya por brindarme sus conocimientos para el desarrollo del presente proyecto de titulación, así como también, a mi asesora Msc. Ana Umaquinga que supo brindarme una asesoría de calidad para la culminación del proyecto, y por último agradecer a los ingenieros de la carrera de Telecomunicaciones que a lo largo de mi vida universitaria compartieron sus conocimientos y me enseñaron de esta linda carrera.

Finalmente quiero agradecer a mi novia, por siempre alentarme en mis momentos más difíciles, por brindarme su ayuda y apoyo incondicional, a mis sobrinos, amigos, compañeros y todas las personas que siempre me brindaron fuerzas y apoyo para la culminación de este trabajo de titulación.

*Carlos Fernando Guerrón Pantoja*

## RESUMEN

Este proyecto de grado se centra en la creación de un "Sistema de Reconocimiento de Emociones a través de la Voz, utilizando Técnicas de Aprendizaje Profundo". Se basa en la Inteligencia Artificial, en particular en el Aprendizaje Supervisado con Redes Neuronales Artificiales, que pueden ser utilizadas para predecir emociones. La necesidad de un sistema de este tipo surge de su potencial uso en psicología para ayudar a detectar patologías de depresión. Para alcanzar los objetivos predeterminados se empleará una metodología en cascada.

Al inicio del proyecto se realiza un estudio bibliográfico, el cual se centra en temas como el funcionamiento del habla y la adquisición de ciertas características de las emociones en la voz. Para ello, se utiliza la transformada Wavelet para extraer características de señales de audio obtenidas de una base de datos de habla emocional mexicana, la cual se utilizó para el desarrollo del sistema. En esta sección también se describen diferentes tipos de redes neuronales y se especifica cuál se eligió para utilizar en el sistema.

Después de esto, se utiliza la metodología KDD (Knowledge Discovery in Database) para el diseño del sistema, donde se emplea una base de datos existente que contiene audios con diferentes emociones. A estos audios se les aplica la transformada Wavelet multinivel, que descompone la señal original en subseñales que presentan características específicas para cada audio. En consecuencia, se aplica el cálculo de energía de cada señal obtenida por la transformada Wavelet para obtener el conjunto de datos de entrenamiento, que luego se normaliza y se genera la arquitectura de la red neuronal LSTM. Los parámetros y la arquitectura de entrenamiento surgen de varias pruebas realizadas y de la comparación del porcentaje de efectividad y error de la red.

Finalmente, se realizan pruebas de rendimiento en pacientes que sufren de patología depresiva, donde se aplica el llamado "test de Beck", que muestra el nivel de depresión que tiene el paciente. En consecuencia, la persona lee un texto donde se registra su voz, y luego se realiza el proceso de extracción de características y reconocimiento de la emoción con la red neuronal que ya ha sido entrenada. El resultado es que el 50% de los pacientes sufren de depresión grave, mientras que el otro 50% sufre de depresión leve, lo que es corroborado por las emociones detectadas por el sistema y el test aplicado.

## ABSTRACT

This degree project focuses on the creation of an "Emotion Recognition System through Voice, using Deep Learning Techniques". It is based on Artificial Intelligence, in particular Supervised Learning with Artificial Neural Networks, which can be used to predict emotions. The need for such a system arises from its potential use in psychology to help detect depressive pathologies. A cascade methodology will be used to achieve the predetermined objectives.

At the beginning of the project a bibliographic study is carried out, which focuses on topics such as the functioning of speech and the acquisition of certain characteristics of emotions in the voice. For this, the Wavelet transform is used to extract features from audio signals obtained from a Mexican emotional speech database, which was used for the development of the system. This section also describes different types of neural networks and specifies which one was chosen to be used in the system.

After this, the KDD (Knowledge Discovery in Database) methodology is used for the design of the system, where an existing database containing audios with different emotions is used. The multilevel Wavelet transform is applied to these audios, which decomposes the original signal into sub-signals that present specific characteristics for each audio. Consequently, the energy calculation of each signal obtained by the Wavelet transform is applied to obtain the training data set, which is then normalized, and the architecture of the LSTM neural network is generated. The training parameters and architecture arise from several tests performed and from the comparison of the percentage of effectiveness and error of the network.

Finally, performance tests are performed on patients suffering from depressive pathology, where the so-called "Beck test" is applied, which shows the level of depression that the patient has. Consequently, the person reads a text where his or her voice is recorded, and

then the process of feature extraction and emotion recognition is performed with the neural network that has already been trained. The result is that 50% of the patients suffer from severe depression, while the other 50% suffer from mild depression, which is corroborated by the emotions detected by the system and the applied test.

# INDICE

# INDICE DE FIGURAS

# INDICE DE TABLAS

# CAPÍTULO 1

Este capítulo ofrece una breve visión general de la motivación del autor para desarrollar el proyecto. También esboza el tema, el problema, los objetivos, el alcance y la justificación que se abordarán a lo largo del estudio, al tiempo que tiene en cuenta los aspectos teóricos necesarios y las limitaciones requeridas para su desarrollo.

## 1.1. Tema

Sistema de reconocimiento de emociones a través de la voz, mediante técnicas de aprendizaje profundo.

## 1.2. Problema

Según (Dirección General de Comunicación Social, UNAM, 2018), alrededor del mundo existe un 10% de personas que no pueden expresar sus emociones de forma corporal o por su estado físico, con lo que conlleva a tener un impedimento al momento de hablar de tener una buena salud mental, ya que al momento de establecer comunicación con otras personas el individuo no presenta seguridad. Las emociones forman parte del ser humano y así cada uno de ellos expresa su estado de ánimo, ya puede ser tristeza, enojo, felicidad, y el estado emocional neutral. Con esto se han venido teniendo varios problemas, el más grave y peligroso es el de la depresión, “los pacientes deprimidos pueden experimentar gran deterioro en su funcionamiento habitual, en su bienestar y también en su calidad de vida” (Orozco & Baldares, 2012).

Según una infografía presentada por el INEC en el año 2015, 2.088 personas en Ecuador fueron atendidas por enfermedades depresivas en los establecimientos de salud del Ecuador, estos datos han ido creciendo con el pasar de los años. El Observatorio Social de Ecuador ha dado a conocer que entre 2014 y 2019 se produjeron en el país 5.300 suicidios (entre dos y tres al día) (Cabezas, 2020), esto debido a que no se detectó el problema en las personas, principalmente es el de la depresión. En la pandemia por el COVID-19, la UEES implemento

una línea de ayuda psicológica, en donde se evidenció que el 62 % de individuos correspondió al sexo femenino y que la principal sintomatología era la depresión (Valcárcel, Santiesteban, & Abad, 2021). En este sentido, el mantener el control sobre el estado de ánimo es extremadamente importante para la salud mental y sobre todo para las personas que van al psicólogo cuando presentan este tipo de emociones, ya que según (Singh & Kaur, 2019): "el habla es una forma eficaz de comunicación para reconocer al hablante y los diferentes tipos de emociones". Según la opinión de dos psicólogos, se tiene un problema al momento de realizar consultas, la falta de elementos externos que ayuden al diagnóstico del paciente. Es aquí donde nace la necesidad de crear un sistema que ayude al reconocimiento de emociones en los pacientes.

El reconocimiento de emociones puede revelar la actitud de los seres humanos a través de sentimientos encontrados en el lenguaje del habla. En este sentido, la detección de las emociones humanas es esencial para el cuidado de la salud personal y el estado mental (Kaur & Pandey, 2018). Por lo que, para el presente proyecto se plantea diseñar un sistema el cual sea capaz de detectar las emociones del paciente que acude al psicólogo para que así el profesional tenga una ayuda extra para realizar su diagnóstico, todo esto gracias a la ayuda del hardware y software que forman parte del sistema, en lo que respecta al software se desarrolla con la ayuda del aprendizaje profundo, el cual la relación hombre-maquina hará que se pueda obtener resultado de las emociones que el paciente desarrolló durante la entrevista que el psicólogo aplica.

Gracias a que con la voz podemos detectar las emociones de los pacientes este dispositivo pretende de una forma poder ayudar a las personas en su salud, los diferentes obstáculos que se presenta en la vida de cada persona hace que se tengan problemas con la salud emocional y así presentar algún tipo de cuadro de depresión severa que llevaría a un posible suicidio, con el diseño del sistema se tendrá en cuenta la interacción de las interfaces

hombre-maquina a partir del reconocimiento y procesamiento de las señales que nos arroja la voz de cada uno de los pacientes.

## 1.3. Objetivos

### *1.3.1. Objetivo General*

Diseñar un sistema de Reconocimiento de Emociones de Voz (REV) mediante técnicas de aprendizaje profundo para la ayuda en el diagnostico de depresión en pacientes que acuden al psicólogo.

### *1.3.2. Objetivos Específicos*

- Realizar un estudio bibliográfico acerca del reconocimiento de emociones a través de la voz, y del aprendizaje profundo.
- Diseñar la arquitectura de aprendizaje profundo con el modelo de descubrimiento de conocimiento en base de datos (KDD), para que el sistema sea capaz de detectar las emociones en los pacientes.
- Implementar el sistema de reconocimiento de emociones y realizar pruebas de funcionamiento.
- Verificar mediante porcentajes de precisión del sistema los resultados obtenidos mediante pruebas de voz de diferentes pacientes

## 1.4. Alcance

El proyecto tiene como finalidad la creación de un sistema que con técnicas de aprendizaje profundo este sea capaz de detectar las emociones en personas que acuden a una consulta con el objetivo de realizar la implementación en el área de atención a pacientes, esto para tener un apoyo extra hacia el psicólogo y mediante la aplicación del criterio profesional que se pone en ejecución, poder brindar un diagnóstico mejorado al paciente.

Se creará una base bibliográfica en donde consten parámetros de las emociones esenciales de las personas y como estas influyen en la personalidad de cada uno de nosotros, también se investigara bases teóricas que ayuden a fundamentar conceptos sobre la anatomía del habla, aquí se verificara los elementos del habla, frecuencia fundamental (pitch) y el procesamiento digital de las señales de la voz, para así poder crear el algoritmo de aprendizaje profundo en Matlab y hacer que el sistema verifique mediante las señales de voz el tipo de emoción que el paciente está sintiendo en el momento de la realización de pruebas.

Mediante un modelo en cascada se realizará la correcta elección de los requerimientos del sistema, donde el proceso del desarrollo del sistema cuentas con las fases sucesivas, como el análisis, en donde se abordarán los conceptos que se manejan en el proyecto. Otra de las fases del proyecto es el diseño y las especificaciones en donde se constata la programación, entrenamiento basado en la metodología de Descubrimiento de Conocimiento en Base de Datos (Knowledge Discovery in Databases: KDD), implementación, pruebas de funcionamiento y al final una verificación de los resultados obtenidos.

En el segmento de diseño se establecerá un modelo KDD, que empieza con la selección de la base de datos que usaremos para que el sistema se entrene, es decir; en la base de datos constaran audios en donde personas actúen las emociones que vamos a establecer (tristeza, enojo, alegría y estado neutral), con esto ya tendremos lo datos de entrenamiento, el siguiente paso en la metodología es la de procesamiento, la voz de la persona es procesada mediante una conversión análogo-digital. Con la ayuda del software Matlab se realiza la transformación de las señales, para dicha transformación se aplicará la denominada Transformada de Wavelet, con esto podremos sacar las características fundamentales de las señales de cada emoción que serán parte de la red neuronal, una vez que se establece la red neuronal el sistema será capaz de detectar las emociones.

Se realizará la implementación del sistema, en este segmento se pondrá en funcionamiento el proyecto, empezando por corroborar el funcionamiento del sistema, luego de esto se establecerá contacto con diversas personas para someter al sistema en pruebas de funcionamiento real, aquí se obtendrán resultados de las mediciones de las emociones para después ser analizadas.

Finalmente se obtendrán los resultados de las pruebas realizadas con el sistema a diversas personas, en donde se clasificará la eficacia de la medición de las emociones, así tendremos el porcentaje en cada una de las emociones medidas y verificaremos cuál de ellas es la de menor porcentaje de exactitud, precisión y sensibilidad para que el administrador del sistema pueda tener en cuenta esos valores y así darle el uso debido a esta herramienta de ayuda.

### 1.5. Justificación

Las emociones son reacciones psicofisiológicas de las personas ante situaciones relevantes desde un punto de vista adaptativo (Rodríguez, Linares, González, & Guadalupe, 2009). Factores como la presencia de la pandemia, diversas crisis económicas y sociales hacen que crezca las enfermedades emocionales, con la realización del proyecto se pretende ayudar a profesionales de la psicología a tener información extra en consulta, adjuntando tecnologías nuevas, en este caso la inteligencia artificial.

La inteligencia artificial (IA), es una tendencia en la actualidad, ya que; varios proyectos se enfocan a usar esta técnica en sus investigaciones y así aportar nuevo conocimiento y estrategias que mejoran la vida de las personas, en este proyecto se pretende estudiar las emociones de las personas basados en la forma de hablar de cada uno de ellos, así para poder establecer un criterio de aprendizaje de maquina se toma como referencia las redes neuronales convolucionales que están inspiradas en las redes neuronales biológicas del cerebro humano. Están constituidas por elementos que se comportan de forma similar a la neurona biológica en sus funciones más comunes. Estos elementos están organizados de una forma

parecida a la que presenta el cerebro humano (Olabe Basogain, 2015). Con esta idea se forma un sistema que cuente con los recursos de aprender de forma autónoma y sea capaz de detectar las emociones en personas, en este caso, personas que acuden al psicólogo, el profesional de la materia tendrá un recurso extra que durante la sesión podrá evaluar el comportamiento emocional del paciente.

La inteligencia artificial, el aprendizaje profundo y el aprendizaje automático son fuentes dominantes de uso para hacer un sistema más inteligente (Mustaqeem & Kwon, 2020). Gracias a estos algoritmos y la teoría que se tiene de ellos, se crea varios aportes en la investigación, tal es el caso del entrenamiento del sistema basado en audio, ya que, se tomaran muestras de voz de diferentes personas en destinos estados de ánimo, lo cual ayudara con la interpretación de las emociones de futuros pacientes que acuden y use este sistema ya entrenado. Todo esto se lo vincula con el hardware que se encarga de obtener los datos y estos resultados se los mostrara en una interfaz gráfica en donde el psicólogo pueda apreciar el resultado de dicho sistema. En el mercado de los circuitos integrados existe una gran variedad de productos, varios tipos de módulos para comunicaciones inalámbricas WIFI y en su corazón está colocado un microcontrolador de altas prestaciones y de un muy bajo costo (Zambrano Galarza, 2021).

La salud mental abarca una amplia gama de actividades directa o indirectamente relacionadas con el componente de bienestar mental (Fernández, Valdespino, Palacios, Guerrero, & Acevedo, 2020). Realizando el sistema se podrá tener datos de las emociones de los pacientes que acuden al psicólogo, siendo ellos los principales beneficiarios, ya que con esto el profesional tendrá un mejor criterio de diagnóstico. Todo esto brindando un aporte al reconocimiento de emociones mediante la voz con redes neuronales.

# CAPÍTULO 2

El objetivo de este capítulo es ofrecer una visión general de las técnicas de aprendizaje profundo utilizadas para el reconocimiento de emociones basado en el habla. Esto incluye una descripción detallada de las bases de datos del habla, la extracción de emociones y las limitaciones que conlleva el proceso de clasificación de emociones

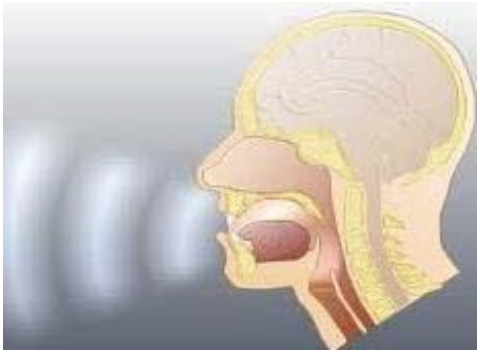
## 2.1. El habla

Es la habilidad para comunicarse verbalmente, además, esta contiene información lingüística que permite detectar palabras o frases, así como características del lenguaje que permiten extraer información del habla. La voz es el sonido producido por la vibración de las cuerdas vocales de los organismos vivos y tiene cualidades como timbre, intensidad y calidad (Jahangir et al., 2021).

El ser humano es capaz de producir diversos sonidos mediante la alteración fonológica, que consiste en convertir la presión del aire en los pulmones en vibraciones audibles al pasar el aire por las cuerdas vocales. Al hablar, las cuerdas vocales actúan como un vibrador, como se muestra en la Figura 1.

**Figura 1.**

*Generación de la voz*



*Nota.* Adaptado de (Sandoval,j 2019)

Las cuerdas vocales tienen su propia frecuencia de resonancia, que determina el timbre de la voz de una persona. Esta frecuencia es la frecuencia natural de oscilación determinada por las características físicas de los pliegues. En los hombres adultos, la longitud de los pliegues

laríngeos suele oscilar entre 17 y 23 mm, mientras que en las mujeres adultas oscila entre 12,5 y 17 mm. Cada pliegue puede extenderse 3-4 mm a lo largo de la laringe. Debido a estas diferencias en las características físicas, la voz masculina suele oscilar entre 110 Hz y la femenina entre 210 Hz. Como resultado, un pequeño chorro de aire es emitido hacia el exterior, produciendo un sonido cuya frecuencia corresponde a la expansión de los pliegues laríngeos (Sandoval, 2019).

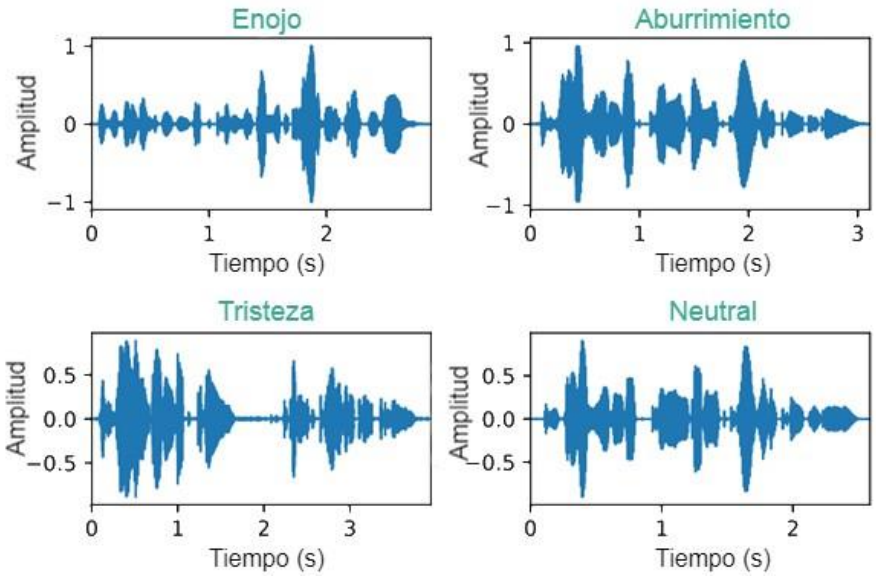
## 2.2. Representaciones de sonido y habla

El proceso de extracción de parámetros acústicos de las expresiones vocales se basa en la idea de las variaciones en el habla, causadas por diferentes estados de excitación o valencia en el hablante, de tal forma que, pueden estimarse mediante diferentes parámetros de onda (Rintala, 2020).

La Figura 2 muestra gráficos de ondas del conjunto de datos EmoDB en alemán (Rintala, 2020). Cada gráfico representa una representación emocional diferente de la frase "Ich will das eben weg bringen und dann mit Karl was trinken gehen", que se traduce como "Solo quiero quitarme eso y luego ir a tomar algo con Karl".

**Figura 2.**

*Emociones retratadas de la base EmoDB sin procesar*.




*Nota.* Adaptado de (Sandoval, 2019).

Las características de una señal del habla son espectrales (el sonido de la voz), prosódicas (la melodía del habla), fonéticas (teléfonos hablados, reducciones y elaboraciones), idiolectales (elección de palabras) y semánticas. En este sentido, las características prosódicas incluyen la Tasa de Cruce por Cero (Zero Crossing Rate: ZCR) y la Energía Media Cuadrática (Root-Mean-Square: RMS).

Existe un conjunto de parámetros acústicos que pueden utilizarse para estimar las variaciones en el habla. Estos Descriptores de Bajo Nivel (Low Level Descriptors: LLD) se pueden ordenar por los siguientes grupos de parámetros: parámetros relacionados con la frecuencia como tono, fluctuación y formante 1; parámetros relacionados con la energía/amplitud como intensidad media del habla, volumen, relación entre armónicos y ruido; parámetros espectrales (equilibrio/forma/dinámica) como relación alfa, índice de Hammar-Berg, diferencia armónica, pendiente espectral y los Coeficientes Cepstrales de las Frecuencias de Mel 1–4 (Mel Frequency Cepstral Coefficients: MFCC) (Rintala, 2020).

Asimismo, existen algunas características temporales, como los picos de tasa de sonoridad y el número de regiones sonoras continuas por segundo (tasa de pseudosílabas). Todos estos LLD se pueden combinar y apilar unos encima de otros en funciones estadísticas (Wang et al., 2022), por ejemplo, se puede calcular la media del tono LLD, es decir, la frecuencia fundamental en una cierta granularidad de escala de tiempo cada 0,1 s, y luego calcular la media de la frecuencia fundamental (en 0,1 s) por cada 0,5 s. Este procesamiento de datos puede dar como resultado un vector de características de alta dimensión, que se puede aplicar para el reconocimiento de emociones (Sandoval, 2019).

La mayoría de los LLD y los métodos de análisis de audio no funcionan a la tasa de muestreo nativa de la señal, sino en pequeños cuadros de las señales, separados por una longitud de salto. Las longitudes de fotogramas y saltos predeterminadas se establecen en 2048 y 512 muestras respectivamente (Rázuri et al., 2015). Esto se relaciona con una frecuencia de

muestreo de 16kHz, que corresponde a fotogramas de 128ms y 32ms, es decir, superpuestos en 96ms (Khalil et al., 2019).

### 2.3. ¿Cómo se obtiene técnicamente la voz de forma digital?

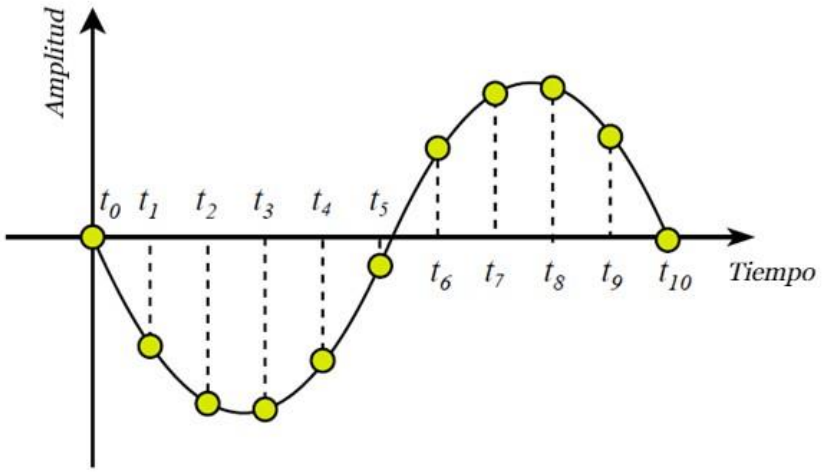
Por muestreo se entiende el proceso de transformación de una señal continua en una secuencia discreta de valores, como se ilustra en la Figura 3. Para convertir una señal analógica en digital, es necesario muestrear a una frecuencia suficientemente alta para no perder información. Para ello, Shannon desarrolló el teorema de muestreo de Nyquist-Shannon (Rintala, 2020), que establece que la frecuencia de muestreo debe ser al menos dos veces la frecuencia máxima de la señal analógica al digitalizar una señal. Esto puede expresarse matemáticamente como

$$f_s \geq 2 * f_c$$

Donde $f_s$ es la tasa de muestreo y $f_c$ es la mayor frecuencia incluida en la señal.

**Figura 3.**

*Muestreo de una señal analógica*




Nota. *Adaptado de (Sandoval, 2019).*

## 2.4. Extracción y modelado de características

Dadas las expresiones emparejadas, la extracción de características busca obtener características que caractericen la emoción en el habla. Las secuencias de características emparejadas resultantes se alinean para obtener una alineación a nivel de fotograma. Es común utilizar representaciones espectrales de baja dimensión a partir de espectros de alta dimensión para el modelado y la manipulación. Las características espectrales comúnmente utilizadas incluyen Coeficientes Cepstrales de Mel (Mel Cepstral Coefficients: MCC), Coeficientes Cepstrales Predictivos Lineales (Linear Predictive Cepstral Coefficients: LPCC) y Frecuencias Espectrales de Línea (Line Spectral Frequencies: LSF) (Zhou et al., 2022).

Como la emoción es compleja con múltiples atributos de señal relacionados con el espectro y la prosodia, tanto los componentes espectrales como los prosódicos necesitan el mismo nivel de atención en la conversión de voz emocional. Por lo general, se consideran varias características prosódicas, como el tono, la energía y la duración. En este sentido, F0 es un componente prosódico esencial que describe la entonación, ya sea lingüística o emocional, en diferentes duraciones que van desde la sílaba hasta el enunciado (Zhou et al., 2022).

Los métodos para modelar las variantes F0 incluyen, estilización y modelado multinivel. En el método de modelado multinivel, la Transformada Continua de Wavelet (Continuous Wavelet Transform: CWT) se utiliza para modelar características prosódicas jerárquicas, como F0 y contorno de energía (Sisman, 2019). Con el análisis CWT, una señal se puede descomponer en componentes de frecuencia y representar con diferentes escalas temporales. CWT ha demostrado ser eficaz para modelar la prosodia del habla y se ha aplicado con éxito en varios marcos de conversión de voz emocional (Zhou et al., 2020).

## 2.5. Transformada Wavelet continua

En el proceso de la transformada wavelet, primero emplea la señal de voz sin procesar como señal de entrada, y luego es indispensable que se explore la base wavelet de Daubechies

(dbN) para dividir la señal en componente de aproximación y detalle respectivamente, donde h representa baja frecuencia dominio y g demuestra dominio de alta frecuencia. Luego, se divide aún más la señal procesada anteriormente en dominio bajo-alto por la misma analogía. Además, es importante seleccionar la función base wavelet adecuada que posea suficiente capacidad para igualar la señal sin procesar en el dominio de tiempo-frecuencia. En este sentido, se consideran múltiples factores para elegir la wavelet de Daubechies (dbN) que presenta Inrid Danbechies. La wavelet dbN tiene características especiales que incluyen simultáneamente mejor ortogonalidad, soporte compacto y mayores órdenes de momento de fuga, con el aumento de los rangos de secuencia para que sea más fuerte para la capacidad de localización en el dominio de la frecuencia (Meng et al., 2021).

La estructura completa de la descomposición del paquete wavelet se muestra en la Figura 4. El propósito general es adquirir una nueva señal de voz de reconstrucción del paquete wavelet en comparación con la señal sin procesar, para lo cual, se define el componente aproximado Φ (t) como d10 (t) y el componente detallado será d11 (t) en la primera capa, donde el subíndice apunta al número de capa de descomposición de ondículas y el superíndice apunta a la ubicación del paquete de ondículas en la capa. En primer lugar, se debe calcular el valor de la base de la función del paquete wavelet que se muestra en la Ecuación (1) para retener más información de dominio temporal y de frecuencia (Meng et al., 2021). Asimismo, las ecuaciones se reescriben como se presenta en la Ecuación (2).

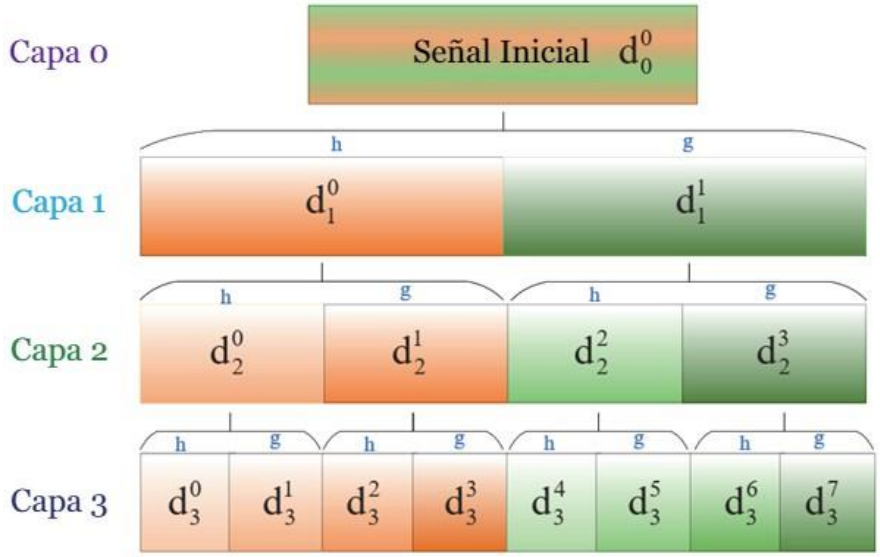
$$d_{L-1}^{2n}(t) = \sum_k h_K \, d_L^n(t-k), d_L^n(t-k) = d_{L-1}^n(2t-k)$$
$$d_{L-1}^{2n+1}(t) = \sum_k g_K \, d_L^n(t-k), d_L^n(t-k) = d_{L-1}^n(2t-k) \quad (1)$$

$$d^{2n}(t) = \sum_k h_K \, d^n(2t-k)$$
$$d^{2n+1}(t) = \sum_k g_K \, d^n(2t-k) \quad (2)$$

En este sentido, hk y gk representan filtros de media banda de paso bajo y paso alto respectivamente, además, se adopta la transformada wavelet diádica que discretiza la escala por series de potencias. Los parámetros de escala son 2i, que denotan el número de capas. Los parámetros de d y k expresan el coeficiente del paquete wavelet y la variable de traducción por separado (Meng et al., 2021).

**Figura 4.**

*Estructura de descomposición de paquetes wavelet*




*Nota.* Adaptado de (Meng et al., 2021)

Luego, se calcula el valor de la transformación del paquete de ondículas que se usa para hacer coincidir la señal de voz sin procesar, como el coeficiente de base de la función del paquete de ondículas que se presenta en la Ecuación (3). Esta última demuestra el significado de adquirir los valores de proyección en cada base de función de paquete de ondículas, a partir de la señal de voz sin procesar de acuerdo con el cálculo del producto interno entre los parámetros anteriores. Cuantos más valores de proyección hay, más porcentajes de información de características son transportados por la señal sin procesar emparejada con la señal de ondícula (Meng et al., 2021).

$$D_{\Phi}^{2n}\left(2^i,k\right) = < f(t), d^{2n}(t) > \frac{1}{\sqrt{2^i}} \sum_t \Phi * \left(\frac{2t-k}{2^i}\right)$$
$$D_{\Phi}^{2n+1}\left(2^i,k\right) = < f(t), d^{2n+1}(t) > \frac{1}{\sqrt{2^i}} \sum_t \Psi * \left(\frac{2t+1-k}{2^i}\right) \quad (3)$$

Donde f (t) representa la señal de voz sin procesar. Por último, se debe generar ocho nuevas señales de voz reconstruidas con tres capas después de explotar el valor de la transformación del paquete wavelet, que se muestra en la Ecuación (4).

$$f_{new}^{*2n} = \sum (D_{\Phi}^{2n} * d^{2n}(t))$$
$$f_{new}^{*2n+1} = \sum (D_{\Psi}^{2n+1} * d^{2n+1}(t)) \quad (4)$$

Por lo tanto, las $f_{news}$ se utilizan como nuevas señales de voz para extraer más mapas de características. Además, es un hecho notable que se debe seleccionar ocho nuevas señales reconstruidas para ejecutar la siguiente extracción de características. Debido a que, si se elige menos de ocho hojas, puede ocurrir que la distribución del dominio de la frecuencia no sea minuciosa y podría llevar a ignorar mucha información de frecuencia útil. Mientras que, si se elige más de ocho hojas, la distribución es tan redundante que podría generar mucha información de frecuencia innecesaria y perder tiempo de entrenamiento. Por lo tanto, es más adecuado para redes modelo explotar tres capas de paquetes wavelet (Meng et al., 2021).

**2.6. Redes Neuronales**

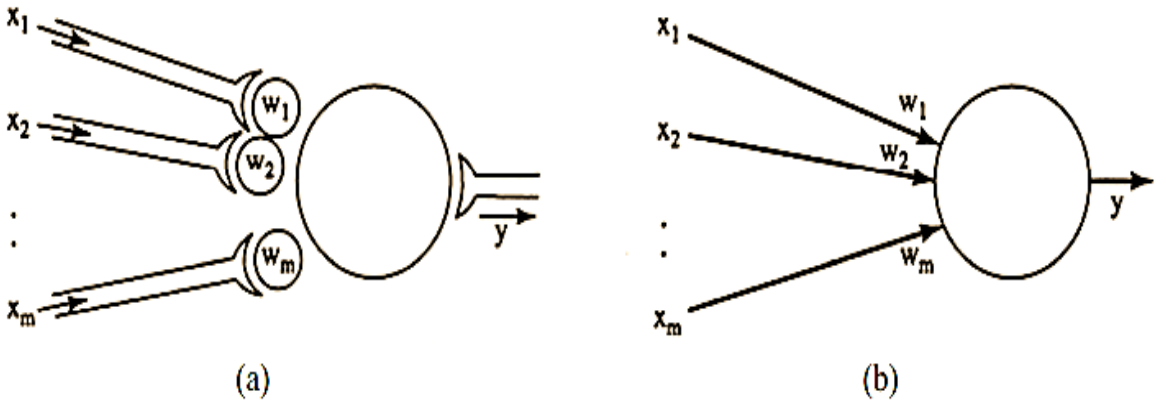
Una Red Neuronal (NN) es un modelo computacional que emula el comportamiento del cerebro humano, compuesto por aproximadamente 1011 unidades conocidas como neuronas. Estas neuronas están interconectadas mediante enlaces, lo que da lugar a más de 1015 conexiones. Por ello, el cerebro se considera la principal fuente de inteligencia, que abarca el aprendizaje, la cognición y la percepción.

Por el contrario, una red neuronal está formada por neuronas artificiales interconectadas, representadas por nodos o vértices, y conexiones representadas por aristas. La neurona artificial, elemento fundamental de una red neuronal, emula a su homóloga natural que se encuentra en el cerebro.

Una neurona artificial es una réplica de una neurona natural, como se representa en la Figura 5. Se compone de entradas (x1, x2, ..., xm) que imitan los niveles de estimulación naturales. El cuerpo de la neurona es estimulado por el producto de sus entradas (xi) y los pesos de las fuerzas sinápticas biológicas (wi).

**Figura 5.**

*Modelo de neurona artificial que se basa en la neurona natural e incorpora la abstracción de red neuronal*




*Nota.* Adaptado de (Toshinori, 2008)

La neurona ejecuta una suma de la multiplicación para i = 1 a m, que se denomina suma ponderada. En este ejercicio intervienen dos vectores, x=[x1,x2,...,xm] e yw=[w1,w2,...,wm], sobre los que se aplica un producto escalar para obtener el valor neto, que se calcula como net=x1w1+x2w2+⋯+xm*wm. Posteriormente, la neurona utiliza una función de activación o transferencia, denotada por y=f(net), para determinar el valor de salida a partir de la matriz resultante (Toshinori, 2008).

### 2.6.1. Inteligencia artificial (IA)

La inteligencia artificial se ocupa del análisis y la automatización del comportamiento inteligente y abarca diversos ámbitos, como el mundo natural, personas, animales y plantas. Por ello, existen varias definiciones válidas de IA, que dependen del área de estudio. Estas definiciones incluyen la capacidad de aprender de la experiencia, adaptarse eficazmente a nuevas situaciones, que la inteligencia de un individuo sea proporcional a sus conocimientos adquiridos y la capacidad de un organismo para resolver problemas novedosos.

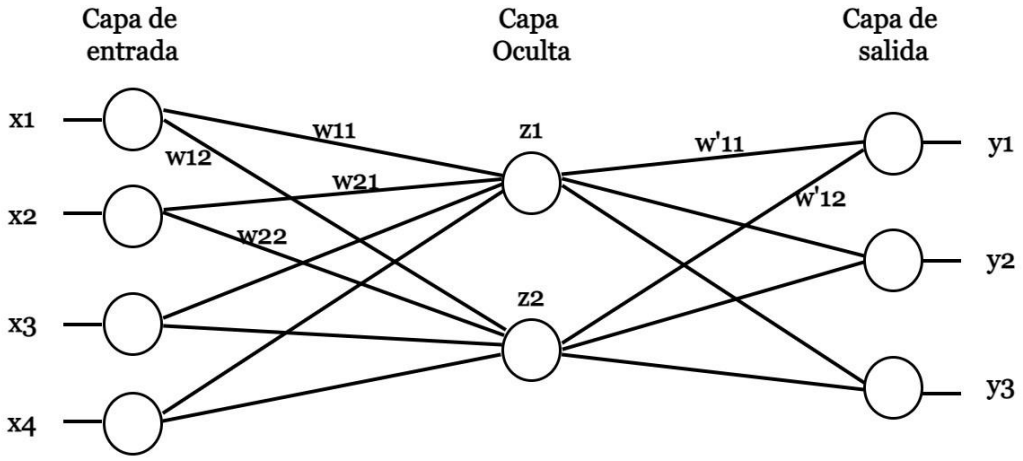
El objetivo de la IA es replicar la inteligencia o cognición humana, sin tener necesariamente en cuenta las emociones y sensaciones de un ser humano. Este objetivo desempeña un papel crucial en el desarrollo de artefactos útiles que satisfagan las necesidades humanas, sin nociones subjetivas o concepciones abstractas de la inteligencia (Chowdhary, 2020).

### 2.6.2. Arquitectura

La forma en que las neuronas están conectadas entre sí se denomina arquitectura de red. En una red neuronal existen varias capas de neuronas: la de entrada, la oculta y la de salida. Un ejemplo de arquitectura en capas es el modelo de retropropagación, ilustrado en la Figura 6, que consta de tres capas compuestas por cuatro neuronas de entrada, dos neuronas ocultas y tres neuronas de salida.

**Figura 6.**

*Modelo backpropagation*




*Nota.* Adaptado de (Calin, 2020)

Normalmente, la arquitectura de una red neuronal consta de una capa de entrada, una capa de salida y un número variable de capas ocultas. No obstante, hay algunas arquitecturas muy utilizadas, como el modelo con una sola capa oculta, o sin capas ocultas. Por otro lado, los modelos con dos o más capas ocultas se utilizan con menos frecuencia (Calin, 2020).

### 2.6.3. Topología de red neuronal

La topología, o disposición de las neuronas en una red neuronal, desempeña un papel fundamental en su aprendizaje y funcionamiento. Esto se debe a que la topología está vinculada a la forma en que se conectan las neuronas. Los mapas autoorganizados son un ejemplo común de topología de aprendizaje no supervisado, que asigna directamente las entradas a un conjunto de categorías o unidades.

Por otro lado, las redes feedforward se componen de tres capas conectadas entre sí. Todos los valores de entrada a la red se conectan a las capas ocultas, mientras que la salida de las neuronas ocultas se conecta a las neuronas de la capa de salida. Estas redes son populares porque teóricamente son capaces de aproximar funciones universales, como la sigmoidea y la lineal gaussiana. En la práctica, las redes neuronales que tienen más capas suelen entrenarse

con modelos de correlación y redes de creencia profunda, ya que el proceso de entrenamiento se simplifica utilizando estos modelos (Miikkulainen, 2017).

### 2.6.4. Tipos de aprendizaje

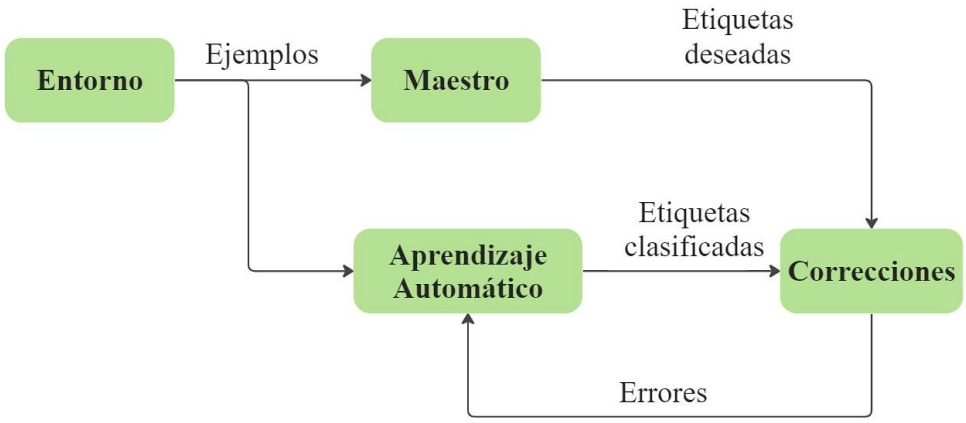
En una red neuronal, el aprendizaje consiste en exponer un conjunto de patrones para que la red los aprenda. Esto se consigue alterando los pesos asociados a las conexiones sinápticas y aplicando reglas de aprendizaje. Existen cuatro tipos de aprendizaje en las redes neuronales: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado y aprendizaje reforzado, que se describen a continuación.

#### *Aprendizaje supervisado*

El aprendizaje supervisado implica dos bloques iniciales con valores diferentes. El primero es la salida objetivo que se asigna inicialmente a cada elemento de formación (Maestro), y el segundo es la salida calculada obtenida a partir del algoritmo de aprendizaje automático, como se muestra en la Figura 7. El objetivo principal de este tipo de aprendizaje es minimizar la diferencia entre la salida objetivo y la salida calculada. El objetivo principal de este tipo de aprendizaje es minimizar la diferencia entre la salida objetivo y la salida calculada. Algunos ejemplos de algoritmos de aprendizaje supervisado son Naive Bayes, árboles de decisión y K-Nearest-Neighbor (KNN), muy utilizados en los ámbitos de la regresión y la clasificación.

**Figura 7.**

*Diagrama consecuente de aprendizaje supervisado*



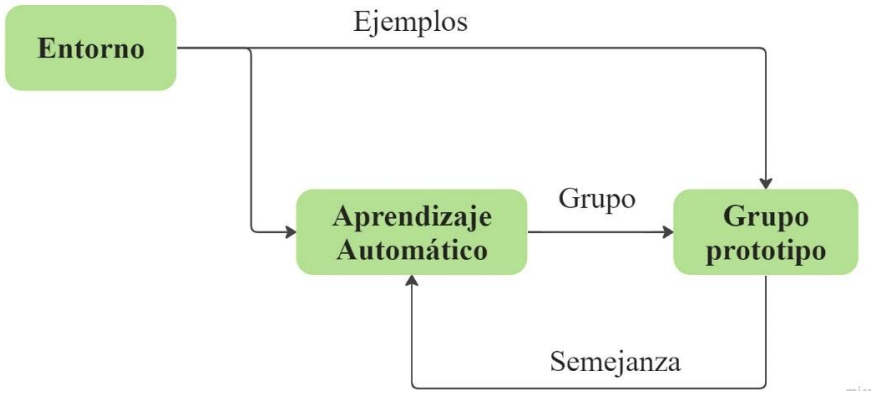

*Nota.* Adaptado de (Miikkulainen, 2017)

***Aprendizaje no supervisado***

El aprendizaje no supervisado implica un bloque de entrada o entorno en el que se introducen los datos iniciales para crear ejemplos de entrenamiento. A diferencia del aprendizaje supervisado, los datos no están etiquetados y los prototipos de clúster se inicializan aleatoriamente, como se muestra en la Figura 8. El proceso de optimización de los prototipos de clúster se realiza en función de las similitudes entre los ejemplos de entrenamiento. El proceso de optimización de los prototipos de clúster se realiza en función de las similitudes entre los ejemplos de entrenamiento. Los algoritmos de aprendizaje no supervisado se utilizan habitualmente en la agrupación de conjuntos de datos (Jo, 2021).

**Figura 8.**

*Diagrama consecuente de aprendizaje no supervisado*



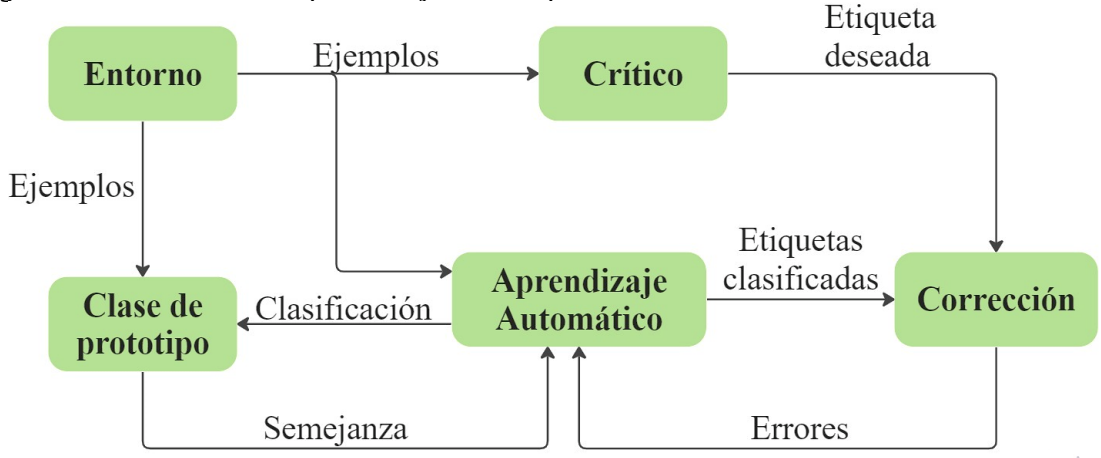

*Nota.* Adaptado de (Jo, 2021)

***Aprendizaje semi supervisado***

El aprendizaje semisupervisado es una combinación de aprendizaje supervisado y no supervisado. En este tipo de aprendizaje, se utilizan ejemplos etiquetados y no etiquetados para entrenar el algoritmo de aprendizaje automático. Los ejemplos etiquetados se utilizan para minimizar el error entre la etiqueta objetivo y la etiqueta calculada, mientras que los ejemplos no etiquetados ayudan a mejorar la precisión del modelo proporcionando información adicional. La figura 9 ilustra el proceso de aprendizaje semisupervisado (Jo, 2021).

**Figura 9.**

*Diagrama consecuente de aprendizaje semi supervisado*



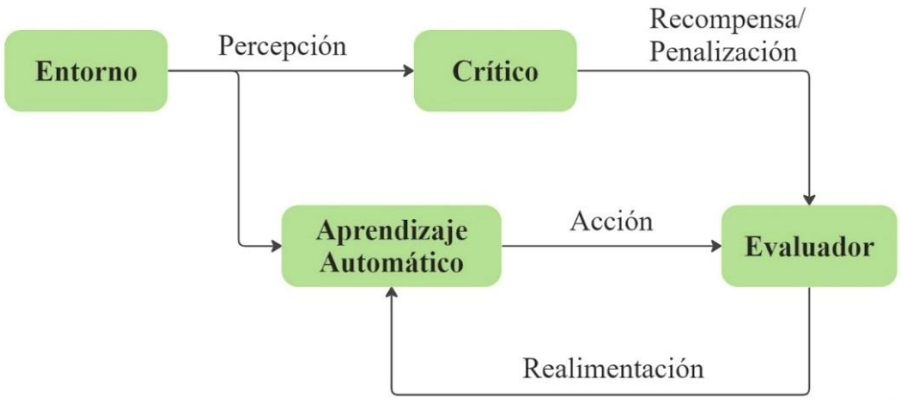

*Nota.* Adaptado de (Jo, 2021)

***Aprendizaje reforzado***

El aprendizaje por refuerzo es un tipo de aprendizaje cuyo objetivo es maximizar la recompensa obtenida de un entorno externo. En este tipo de aprendizaje, el algoritmo de aprendizaje automático recibe información del entorno y genera una acción de salida. La recompensa y la penalización se obtienen del entorno como retroalimentación, y el algoritmo actualiza sus parámetros para maximizar la recompensa y evitar la penalización. El proceso de aprendizaje es iterativo y el algoritmo aprende por ensayo y error, ajustando sus parámetros en función de la información recibida del entorno. El aprendizaje por refuerzo se utiliza habitualmente en ámbitos como la robótica, los juegos y los sistemas de control. La Figura 10 ilustra el proceso de aprendizaje por refuerzo (Jo, 2021).

**Figura 10.**

*Diagrama consecuente de aprendizaje reforzado*




*Nota.* Adaptado de (Jo, 2021)

## 2.7. Reconocimiento de emociones y las técnicas de aprendizaje profundo

El reconocimiento de emociones del habla es un aspecto importante para la Interacción Humano-Computadora (Human Computer Interaction: HCI) (Schuller, 2018). Estos sistemas facilitan la comunicación natural con las máquinas mediante la interacción de voz directa, lo que permite comprender el contenido verbal y facilitar la reacción de las personas (Nassif et al., 2019). Algunas aplicaciones incluyen sistemas de diálogo en varios idiomas, como

conversaciones en centros de llamadas, sistemas de conducción de vehículos a bordo y utilización de patrones de emoción del habla en aplicaciones médicas (Rázuri et al., 2015). No obstante, existen muchos problemas en los sistemas de HCI que aún deben abordarse adecuadamente. Por lo tanto, se requieren esfuerzos para resolver estos problemas de manera efectiva y lograr un mejor reconocimiento de las emociones por parte de las máquinas (Rázuri et al., 2015).

Determinar el estado emocional de los seres humanos es una tarea idiosincrásica que puede usarse como estándar para cualquier modelo de reconocimiento de emociones (Khalil et al., 2019). Entre los numerosos modelos utilizados para la categorización de las emociones, el enfoque emocional discreto se considera fundamental. Este último emplea diversas emociones como la ira, el aburrimiento, el asco, la sorpresa, el miedo, la alegría, la felicidad, la neutralidad y la tristeza. Por otro lado, existen modelos que emplean el espacio continuo tridimensional con parámetros como excitación, valencia y potencia.

El enfoque para el Reconocimiento de Emociones del Habla (Speech Emotion Recognition: SER) comprende principalmente dos fases conocidas como la extracción de características y fase de clasificación de características. En el campo del procesamiento del habla, los investigadores han derivado varias características, como excitación basada en fuentes, propiedades prosódicas, factores de tracción vocal y otras singularidades híbridas. La segunda fase incluye la clasificación de características utilizando clasificadores lineales y no lineales (Hansen et al., 2022).

Los clasificadores lineales más utilizados para el reconocimiento de emociones incluyen las Redes Bayesianas (Bayesianas Network: BN), el Principio de Máxima Verosimilitud (Maximum Likelihood Principle: MLP) y la Máquina de Vectores de Soporte (Support Vector Machine: SVM), donde, la señal de voz se considera no estacionaria. Por lo tanto, se considera que los clasificadores no lineales funcionan de manera efectiva para SER.

En este sentido, existe una gran cantidad de clasificadores no lineales disponibles para SER, incluidos el Modelo de Mezcla Gaussiana (Gaussian Mixture Model: GMM) y el Modelo Oculto de Markov (Hidden Markov Model: HMM) (Dileep & Sekhar, 2014). Estos se utilizan para la clasificación de la información que se deriva de las características de nivel básico. Las funciones basadas en energía, como los coeficientes de predicción lineal, los coeficientes dinámicos del espectro de energía de Mel, los coeficientes de cepstrum de frecuencia de Mel y los coeficientes de cepstrum de predicción lineal perceptual se utilizan para el reconocimiento efectivo de emociones a partir del habla. Por otro lado, los clasificadores, como el vecino más cercano, el análisis de componentes principales y los árboles de decisión, también se aplican para el reconocimiento de emociones (Dileep & Sekhar, 2014).

En contraste, las técnicas de aprendizaje profundo son un campo de investigación emergente en el aprendizaje automático y ha ganado más atención en los últimos años (Deng & Yu, 2013). Las técnicas de aprendizaje profundo para SER tienen varias ventajas sobre los métodos tradicionales, lo que incluye su capacidad para detectar estructuras y características complejas sin necesidad de extracción, así como, el ajuste manual de características; tendencia hacia la extracción de características de bajo nivel de los datos sin procesar y la capacidad para tratar con datos no etiquetados (Deng & Yu, 2013).

Por otro lado, las Redes Neuronales Profundas (Deep Neural Networks: DNN) se basan en estructuras de avance compuestas por una o más capas ocultas subyacentes entre entradas y salidas. Las arquitecturas de avance, como las redes neuronales profundas y las Redes Neuronales Convolucionales (Convolucional Neural Networks: CNN), brindan resultados eficientes para el procesamiento de imágenes y videos. Por otro lado, las arquitecturas recurrentes como las Redes Neuronales Recurrentes (Recurrent Neural Networks: RNN), la Memoria a Largo y Corto Plazo (Long Short Term Memory: LSTM) y el Procesamiento del Lenguaje Natural (Natural Language Processing: NLP) son muy efectivas en la clasificación

basada en el habla. Además de su forma efectiva de clasificación, estos modelos tienen algunas limitaciones. Por ejemplo, el aspecto positivo de las CNN es aprender características de datos de entrada de alta dimensión, sin embargo, estas también aprenden características de pequeñas variaciones y distorsiones, por lo que, se requiere una gran capacidad de almacenamiento. De manera similar, los RNN basados en LSTM pueden manejar datos de entrada variable y modelar datos de texto secuencial de largo alcance (Schmidhuber, 2015).

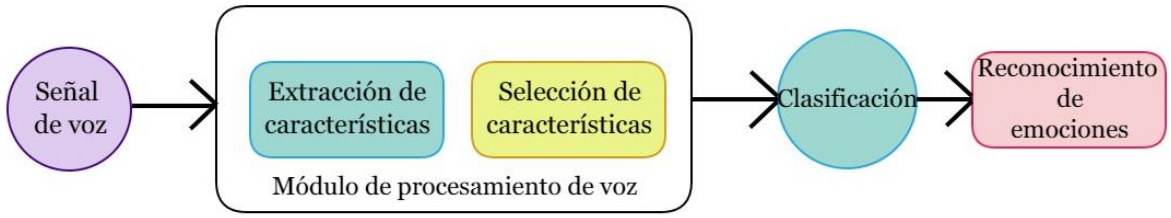
### 2.7.1. Técnicas tradicionales para el SER

Los sistemas de reconocimiento de emociones basados en voz digitalizada se componen de tres partes fundamentales: preprocesamiento de señales, extracción de características y clasificación (Velasco et al., 2022). El preprocesamiento acústico como la eliminación de ruido, así como, la segmentación se llevan a cabo para determinar las unidades significativas de la señal (Anagnostopoulos et al., 2015).

La extracción de características se utiliza para identificar las características relevantes disponibles en la señal. Asimismo, los clasificadores llevan a cabo el mapeo de los vectores de características extraídos a las emociones relevantes. La Figura 11 muestra un sistema simplificado utilizado para el reconocimiento de emociones basado en el habla.

**Figura 11.**

*Sistema tradicional de reconocimiento de emociones del habla*




*Nota.* Adaptado de (Nassif et al., 2019)

En la primera etapa del procesamiento de señales se lleva a cabo la mejora del habla donde se eliminan los componentes ruidosos. La segunda etapa consta de dos partes, la extracción de características y la selección de características. Las características requeridas se extraen de la señal de voz preprocesada y la selección se realiza a partir de las características extraídas. En este sentido, la extracción y selección de características se basa en el análisis de señales de voz en los dominios de tiempo y frecuencia. Durante la tercera etapa, se utilizan varios clasificadores como el algoritmo de Mezclas Gaussianas GMM y los Modelos Ocultos de Markov HMM para la clasificación de características. Por último, en función de la clasificación de características, se reconocen diferentes emociones.

### 2.7.2. Preprocesamiento, extracción y selección de características en SER

Los datos de entrada recopilados para el reconocimiento de emociones a menudo se corrompen por el ruido durante la fase de captura. Debido a estas deficiencias, la extracción y clasificación de características se vuelven menos precisas. Esto significa que la mejora de los datos de entrada es un paso crítico en los sistemas de detección y reconocimiento de emociones. En esta etapa de preprocesamiento se mantiene la discriminación emocional, mientras que, se elimina la variación del hablante y la grabación (Wang et al., 2022).

La señal de voz después de la mejora se caracteriza en unidades significativas llamadas segmentos. Las características relevantes se extraen y clasifican en varias categorías según la información extraída. Un tipo de caracterización es la clasificación a corto plazo basada en características de período corto, como la energía, los formantes y el tono (Aouani & Ayed, 2020). Por otro lado, la media y la desviación estándar son dos de las características esenciales en la clasificación a largo plazo. Entre las características prosódicas, la intensidad, el tono, la velocidad de las palabras habladas y la varianza suelen ser importantes para identificar varios

tipos de emociones a partir de la señal del habla de entrada. Algunas de las características basadas en las emociones acústicas del habla se presentan en la Tabla 1.

**Tabla 1**

*Variaciones acústicas en función de las emociones*

| Emoción | Tono | Intensidad | Velocidad de habla | Calidad de voz |
|---|---|---|---|---|
| **Ira** | abrupto y estrés | mucho más alto | ligeramente más rápido | agitado |
| **Disgusto** | amplio, inflexiones hacia abajo | inferior | mucho más rápido | gruñón |
| **Miedo** | ancho, normal | más bajo | mucho más rápido | voz irregular |
| **Felicidad** | inflexiones mucho más amplias y ascendentes | más alto | más rápido/ más lento | tono entrecortado y a todo volumen |
| **Alegría** | medio alto, rango amplio | más alto | más rápido | velado, timbre a todo volumen |
| **Tristeza** | ligeramente más estrecho | inflexiones hacia abajo | más bajo | resonante |

*Fuente*: (Nassif et al., 2019)

### 2.6.3. Medidas para acústica en SER

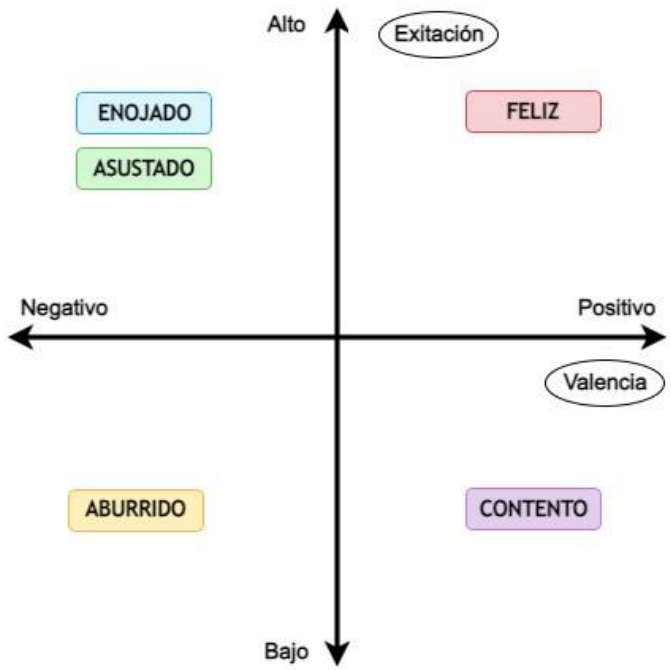
La disponibilidad de información de las emociones está cifrada en todos los aspectos del lenguaje y sus variaciones. Los parámetros vocales y su relación con el reconocimiento de emociones se encuentran entre los temas más investigados en este campo. Con frecuencia se consideran parámetros como la intensidad, el tono y la velocidad de las palabras habladas y la calidad de la voz. Asimismo, se considera una visión directa de la emoción, donde, las emociones existen como categorías discretas. Estas emociones discretas tienen relaciones relativamente claras con los parámetros acústicos, por ejemplo, la intensidad y el tono se correlacionan con la activación, de modo que, el valor de la intensidad aumenta junto con el

tono alto y disminuye con el tono bajo (Yamamoto et al., 2020). Los factores que afectan el mapeo de las variables acústicas a la emoción incluyen, el actuar del hablante, variaciones altas del hablante y el estado de ánimo o la personalidad del individuo (Chen et al., 2020).

En la HCI, las emociones suelen ser espontáneas y no discretas prototípicas, de tal forma que, se expresan débilmente, se mezclan y son difíciles de distinguir entre sí. En la literatura, las declaraciones emocionales se denominan positivas y negativas en función de las emociones expresadas por un individuo. No obstante, algunos experimentos muestran que las emociones actuadas basadas en el oyente son mucho más fuertes y precisas que las emociones naturales, lo que puede sugerir que los actores exageran la expresión de las emociones. En este sentido, las emociones fundamentales pueden ser descritas por áreas dentro del espacio definido por los ejes de excitación y valencia como se aprecia en la Figura 12. La excitación representa la intensidad de la calma, mientras que, la valencia representa el efecto de positividad y negatividad en las emociones (Mauchand & Pell, 2021).

**Figura 12.**

*Espacio emocional bidimensional*




*Nota.* Adaptado de (Nassif et al., 2019)

### 2.7.4. Clasificación de características en SER

En los estudios que proponen un sistema SER, la justificación para elegir un clasificador en particular para la tarea del habla específica, a menudo, no se menciona (Zhu-Zhou et al., 2022). Por lo que, los clasificadores se seleccionan según una regla o evaluación empírica de algunos indicadores.

En contraste, los clasificadores de reconocimiento de patrones utilizados para el SER se pueden clasificar en dos tipos, que son, clasificadores lineales y clasificadores no lineales. Los clasificadores lineales realizan la clasificación en función de las características de los objetos con una disposición lineal de varios objetos (Maji et al., 2022). Estos objetos se evalúan principalmente en forma de una matriz denominada vector de características. Por el contrario, los clasificadores no lineales se utilizan para la caracterización de objetos al desarrollar la combinación ponderada no lineal de dichos objetos. En la Tabla 2 se muestran algunos clasificadores tradicionales lineales y no lineales utilizados para SER.

**Tabla 2**

*Clasificadores lineales y no lineales para SER.*

| Clasificadores | Lineal/no lineal |
|---|---|
| Clasificador Bayesiano | Lineal |
| Clasificador K-vecino más cercano | Lineal |
| Clasificador basado en el Modelo de Mezcla Gaussiana (GMM) | No lineal |
| Clasificador basado en Modelo Oculto de Márkov (HMM) | No lineal |
| Clasificador basado en el análisis de Componentes Principales (PCA) | Lineal/no lineal |
| Clasificador basado en la Máquina Vectorial de Soporte (SVM) | Lineal/no lineal |
| Clasificador basado en las Máquinas de Aprendizaje Extremo (ELM) | Lineal/no lineal |

*Fuente*: (Zhu-Zhou et al., 2022) (Khalil et al., 2019)

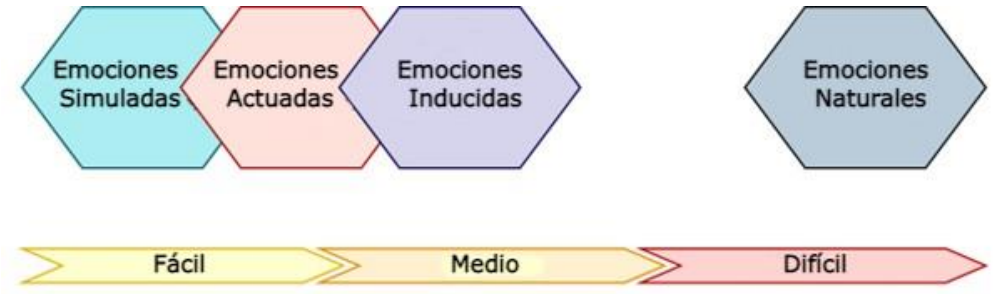
### 2.6.5. Conjuntos de datos empleadas para SER

Las bases de datos emocionales del habla se emplean en una variedad de investigaciones (Sultana et al., 2021; Swain et al., 2018). En este sentido, la calidad de las bases de datos utilizadas y el rendimiento alcanzado son los factores más importantes en la evaluación del reconocimiento de emociones (Sri Lalitha et al., 2021; Valstar et al., 2016). Asimismo, los métodos disponibles y los objetivos en la recopilación de bases de datos de voz varían según la motivación para el desarrollo de sistemas de voz. En la Tabla 3 se proporciona las características de algunas bases de datos de discurso emocional disponibles de forma libre.

Para el desarrollo de sistemas SER, las bases de datos del habla se clasifican en tres tipos, simulada, inducida y natural. Asimismo, la categorización de las bases de datos también se puede describir de acuerdo con el diagrama continuo que se muestra en la Figura 13.

**Figura 13.**

*Bases de datos de emociones y nivel de dificultad*




*Nota.* Adaptado de (Nassif et al., 2019)

**Base de datos simulada:** en estas, los datos del habla son registradas por intérpretes bien entrenados y experimentados (Zhou et al., 2022). Entre todas las bases de datos, esta se considera la forma más sencilla de obtener el conjunto de datos basado en el habla de varias emociones. Además, se considera que casi el 60% de las bases de datos de voz son recopiladas por esta técnica.

**Base de datos inducida:** en este tipo se recoge el conjunto emocional creando una situación emocional artificial (Caschera et al., 2022). Esto se realiza sin el conocimiento del

ejecutante u orador. En comparación con la base de datos basada en actores, esta es una base de datos más naturalista. Sin embargo, puede aplicarse una cuestión de ética, porque, el hablante debe saber que ha sido grabado para actividades basadas en la investigación.

**Base de datos natural:** estas son más realistas, no obstante, son difíciles de obtener debido a la dificultad de reconocimiento (Singh et al., 2022). Las bases de datos del habla emocional natural se registran a partir de la conversación del público en general, las conversaciones del centro de llamadas y terapias. En la Tabla 3 se muestran las bases de datos.

**Tabla 3**

*Bases de datos de discurso emocional disponibles de forma libre.*

| Nro. | Base de datos | Lenguaje | Emociones | Tamaño | Fuente | Tipo de acceso |
|---|---|---|---|---|---|---|
| 1 | Base de datos emocional de Berlín | Alemán | Alegría, Tristeza Aburrimiento Neutralidad Disgusto, Ira Neutralidad | 800 declaraciones | Actores profesionales | Pública y gratuita |
| 2 | Bases de datos emocionales danesas | Danés | Tristeza, Neutralidad Sorpresa Alegría, Ira | 4 actores con 5 emociones | Actores no profesionales | Licencia gratuita |
| 3 | Captura de movimiento diádica emocional interactiva | Inglés | Alegría, Ira Tristeza Frustración Sorpresa, Miedo Asco, Excitación Estado neutral | 10 actores (5 mujeres y 5 hombres) | Actores profesionales | Licencia gratuita |
| 4 | INTERFACE05 | Inglés Español Francés Esloveno | Neutro, Asco Miedo, Alegría Tristeza | Inglés 186 España 184 Francés 175 Esloveno 190 | Actores | Licencia de paga |
| 5 | Discurso emocional y transcripciones LDC | Inglés | Desesperación, Tristeza Neutralidad, Interés, Alegría, Pánico, Ira, Vergüenza Desprecio, Júbilo, Orgullo, Ira | 7 actores con 15 emociones y 10 enunciados | Actores profesionales | Licencia de paga |

*Fuente:* (Zhu-Zhou et al., 2022) (Khalil et al., 2019)

En las primeras investigaciones serias sobre el reconocimiento de emociones basado en el habla, los investigadores comenzaban con bases de datos actuadas y luego pasaban a bases de datos realistas (Martin et al., 2006). Por ejemplo, en cuanto a las bases de datos actuadas se refiere, las bases de datos más utilizadas son la base de datos del habla emocional de Berlín (EmoDB) y la base de datos del habla emocional danesa, que contiene las voces grabadas de 10 artistas. Esto incluye a 4 personas para la prueba, a quienes se les pidió que pronunciaran varias oraciones en 5 estados emocionales diferentes. Los datos comprenden la emoción German-Aibo y los datos Smart-Kom, donde las voces de los actores se graban en un laboratorio. Además, se utilizan las conversaciones del centro de llamadas en un entorno completamente realista a partir de grabaciones en vivo (Burkhardt et al., 2005).

En contraste, existe una gran variación entre las bases de datos, el número de emociones reconocidas, el número de ejecutantes, propósito y metodología. Las bases de datos emocionales del habla se emplean en estudios psicológicos para conocer el comportamiento del paciente, así como, en situaciones donde se desea la automatización en el reconocimiento de emociones (Zloteanu & Krumhuber, 2021). No obstante, el sistema se vuelve complejo y el reconocimiento de emociones es difícil de lograr cuando se emplean datos en tiempo real (Khalil et al., 2019).

### 2.8. Necesidades de técnicas de aprendizaje profundo para SER

El procesamiento del habla funciona de manera sencilla en una señal de audio (LeCun et al., 2015). Es así como, se considera significativo y necesario para diversas aplicaciones basadas en el habla, como SER, eliminación de ruido del habla y clasificación de música (G. Costantini et al., 2022). Con los avances recientes, SER ha ganado mucha importancia. Sin embargo, todavía requiere metodologías precisas para imitar el comportamiento humano para que sea posible la interacción con los seres humanos (Byun & Lee, 2021).

Un sistema SER se compone de varios componentes que incluyen, selección, extracción de características, clasificación de características, modelado acústico, reconocimiento por unidad y modelado basado en lenguaje. Los sistemas SER tradicionales suelen incorporar varios modelos de clasificación, como GMM y HMM (Moine et al., 2021). Los GMM se utilizan para ilustrar las características acústicas de las unidades de sonido, mientras que, los HMM se utilizan para tratar las variaciones temporales que ocurren en las señales del habla.

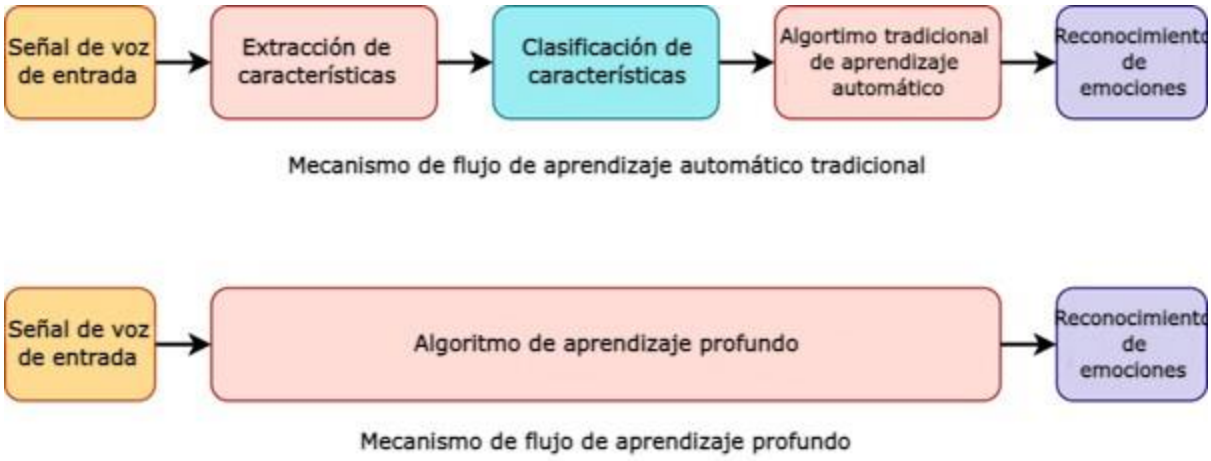
Los métodos de aprendizaje profundo se componen de varios componentes no lineales que realizan cálculos en paralelo (LeCun et al., 2015). Sin embargo, estos métodos deben estructurarse con capas más profundas de arquitectura para superar las limitaciones de otras técnicas. Por ejemplo, las técnicas de aprendizaje profundo como Máquina Profunda de Boltzmann (Deep Boltzmann Machine: DBM), Red Neuronal Recurrente (Recurrent Neural Network: RNN), Red Neuronal Recursiva (Recursive Neural Network: RNN), Red de Creencias Profundas (Deep Belief Network: DBN), Redes Neuronales Convolucionales (Convolutional Neural Networks: CNN) y Codificador Automático (Auto Encoder: AE) se consideran una algunas de las técnicas fundamentales de aprendizaje profundo utilizadas para SER, que mejoran significativamente el rendimiento general del sistema (LeCun et al., 2015).

Por otro lado, el aprendizaje profundo es un campo de investigación emergente en el aprendizaje automático y ha ganado mucha atención en los últimos años (Niu et al., 2017). Es así como, algunos investigadores han empleado Redes Neuronales Profundas para entrenar sus respectivos modelos para SER.

En la Figura 14 se muestra la diferencia entre el flujo de aprendizaje automático tradicional y los mecanismos de flujo de aprendizaje profundo para SER.

**Figura 14.**

*Flujo de aprendizaje automático tradicional frente a flujo de aprendizaje profundo*




*Nota.* Adaptado de (Mahony et al., 2019)

Asimismo, en la Tabla 4 se presenta un análisis comparativo detallado de los algoritmos tradicionales con aprendizaje profundo, es decir, el algoritmo de Red Neural Convolucional Profunda (Deep Convolutional Neural Network: DCNN) en el contexto de la medición de varias emociones utilizando conjuntos de datos IEMOCAP, Emo-DB y SAVEE. En este sentido, la DCNN reconoce emociones como la felicidad, ira y tristeza (Schmidhuber, 2015).

**Tabla 4**

*Análisis comparativo de diferentes clasificadores en SER.*

| **Algoritmo** | **Estado** | | |
|---|---|---|---|
| | **Ira (%)** | **Feliz (%)** | **Triste (%)** |
| k-vecino más cercano | 93 | 55 | 77 |
| Análisis discriminante lineal | 68 | 49 | 72 |
| Máquina de soporte de vectores | 74 | 70 | 93 |
| Análisis discriminante regularizado | 83 | 73 | 97 |
| Red neuronal convolucional profunda | 99 | 99 | 96 |

*Fuente*: (Schmidhuber, 2015) (Nassif et al., 2019)

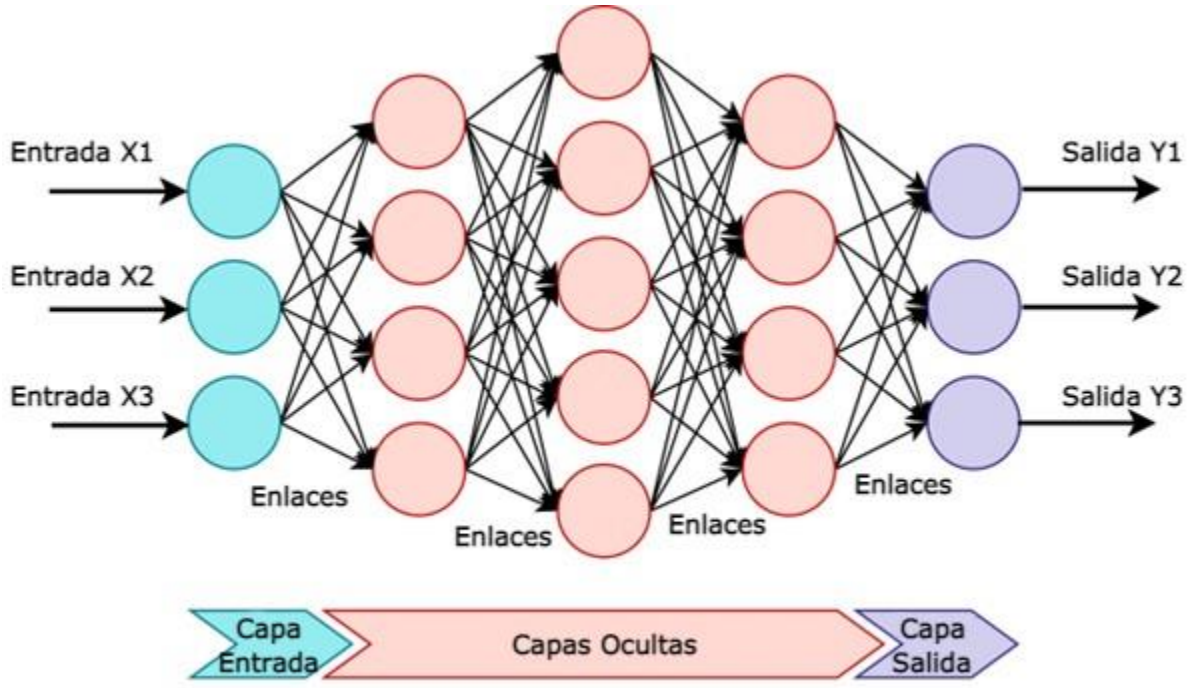
## 2.9. Técnica de aprendizaje profundo para SER

El aprendizaje profundo se deriva de la familia del aprendizaje automático, que es una técnica de aprendizaje más amplia para la representación de datos, como las emociones (Sevilla et al., 2022). Este aprendizaje profundo puede ser no supervisado, semi-supervisado o totalmente supervisado. La Figura 15 ilustra una arquitectura genérica por capas para DNN.

El aprendizaje profundo es un área de investigación de rápido crecimiento debido a su estructura de múltiples capas y la entrega eficiente de resultados. Estas áreas de investigación incluyen reconocimiento de emociones del habla, reconocimiento de voz e imágenes, procesamiento del lenguaje natural y reconocimiento de patrones (Guo, 2022). En este sentido, los algoritmos empleados para los sistemas SER son: máquina de Boltzman Profundo; red de creencia profunda; red Neuronal Convolucional; red neuronal recurrente; red neuronal recursiva, codificador automático y memoria a corto plazo largo.

**Figura 15.**

*Arquitectura genérica de redes neuronales profundas (DNN) por capas*




*Nota.* Adaptado de (Sevilla et al., 2022)

### 2.9.1. Redes neuronales LSTM

La red neuronal con Memoria a Corto y Largo Plazo (Long Short-Term Memory: LSTM) es un tipo diferente de estructura RNN. Esta estructura permite descubrir patrones largos y cortos en los datos, mientras, elimina el problema de la desaparición del gradiente mediante el entrenamiento de RNN. LSTM se ha probado en varias aplicaciones (Jha, 2022) y parece tener un curso muy prometedor en el campo del modelado del lenguaje (Fernandes & Mannepalli, 2021).

El gradiente de fuga parece ser problemático durante el entrenamiento de RNN como se muestra en (Jha, 2022). Esto llevó a los autores a rediseñar la unidad de red, que en LSTM se llama celda. En la Figura 16 se observa que cada celda LSTM contiene puertas que determinan cuándo la entrada es lo suficientemente significativa para recordar, cuándo debe continuar recordando u olvidar el valor y cuándo debe generar el valor. Las celdas así diseñadas pueden interpretarse como una memoria diferenciable (Aggarwal et al., 2022).

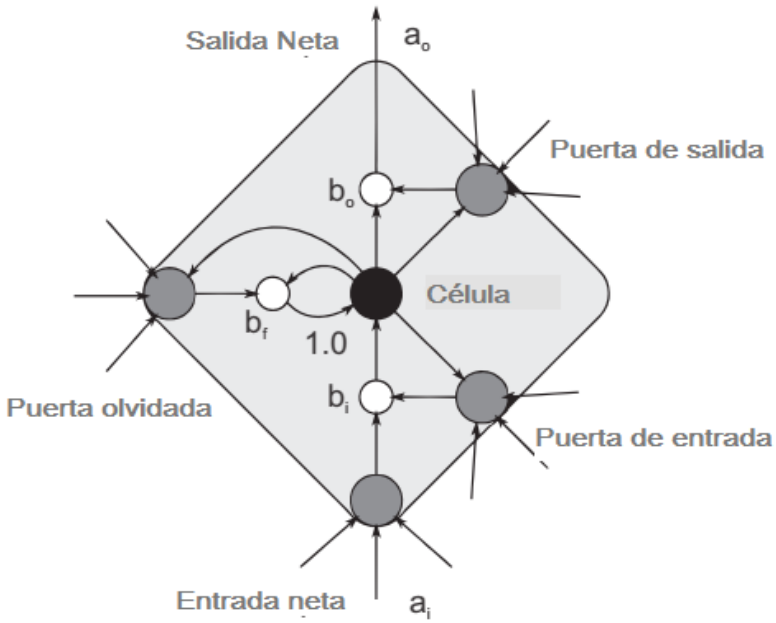
#### Topología LSTM

Una red neuronal típica consiste en la activación de entrada que se transforma en activación de salida con función de activación (generalmente sigmoidal). No obstante, la celda LSTM proporciona esto de manera más integral, donde, las tres entradas de celda llamadas puertas determinan cuándo se permite que los valores entren o salgan de la memoria del bloque que se mira en la Figura 16. En primer lugar, la función de activación se aplica a todas las puertas (Li et al., 2021). Cuando la puerta de entrada genera un valor cercano a cero, pone a cero el valor de la entrada neta, bloqueando efectivamente que ese valor ingrese a la siguiente capa. Cuando la puerta de olvido genera un valor cercano a cero, el bloque olvidará efectivamente cualquier valor que esté recordando. La puerta de salida determina cuándo la unidad debe generar el valor en su memoria. Según el tipo de LSTM, las realizaciones pueden

diferir ligeramente, es decir, introducir algunas modificaciones y mejoras, pero los principios principales son los mismos.

**Figura 16.**

*Celda de memoria LSTM con puertas*




*Nota.* Adaptado de (Brown, 2013)

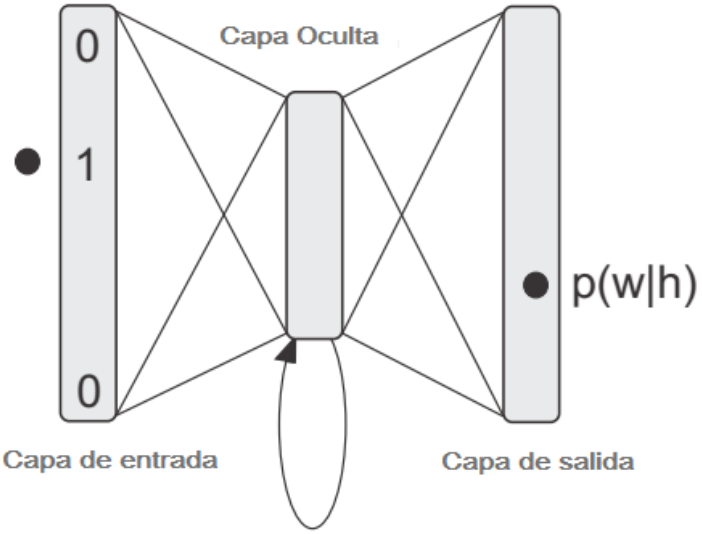
Debido a esta topología específica de LSTM, especialmente debido a un flujo de error constante, la propagación hacia atrás regular podría ser eficaz para entrenar una celda LSTM, puesto que, podría recordar valores durante mucho tiempo. Por otro lado, la LSTM también puede ser entrenado por estrategias de evolución o algoritmos genéticos en aplicaciones de aprendizaje por refuerzo (Li et al., 2021).

**Modelo de lenguaje LSTM**

El LSTM se introdujo con éxito en el campo del modelado del lenguaje, donde, la topología del modelo que se aprecia en la Figura 17 es similar a los modelos de lenguaje RNN comunes que se basan en los siguientes principios: el vector de entrada se codifica en palabras como codificación 1 de N; existe una función softmax utilizada en la capa de salida para producir probabilidades normalizadas; y se utiliza la entropía cruzada como criterio de entrenamiento.

**Figura 17.**

*Arquitectura LMST de red neuronal*




*Nota.* Adaptado de (Brown, 2013)

## 2.10. Redes neuronales en el área de la medicina

En los últimos años, el uso de redes neuronales para abordar problemas complejos se ha hecho cada vez más popular, lo que ha dado lugar a numerosos estudios que exploran la aplicación de redes neuronales artificiales en el ámbito médico. Estos estudios investigan el complejo proceso de descubrimiento de fármacos con diversos fines, como el reconocimiento de patrones, la clasificación, la predicción, el análisis de datos, el pronóstico y el diagnóstico médicos, el control de sistemas de administración de fármacos, el diseño de fármacos y los estudios de estructura-actividad. Además, hay muchas investigaciones que destacan los beneficios para la salud del uso de redes neuronales artificiales en funciones clínicas, como el diagnóstico, el pronóstico, el análisis de supervivencia, los cuidados intensivos y la medicina cardiovascular (Patel & Goyal, 2008).

## 2.11. Detección y diagnóstico de la depresión

La depresión es definida por la Asociación Americana de Psiquiatría como un trastorno mental común que causa tristeza y pérdida de interés en actividades que antes disfrutaba el individuo (Manoret et al., 2021). Se distingue de la tristeza experimentada regularmente como

parte de la vida por su perturbación emocional grave y preocupantemente larga. La depresión muestra síntomas que duran más de dos semanas y tienen fuertes efectos tanto en las funcionalidades como en los comportamientos del paciente.

Los pacientes con depresión también sufren pensamientos de ser inútiles, lo que puede conducir a daños autoinfligidos o incluso al suicidio. Muchos factores de riesgo desempeñan un papel en la causa de la depresión, como el entorno, los factores socioeconómicos y los eventos adversos de la vida (desempleo, eventos traumáticos) (Manoret et al., 2021). Se estima que 1 de cada 15 adultos se ve afectado por la depresión anualmente, y 1 de cada 6 personas experimentará depresión en algún momento de su vida. Sin embargo, existen múltiples métodos efectivos para tratar la depresión que van desde la medicación antidepresiva hasta la psicoterapia interpersonal (OMS, 2021).

Convencionalmente, los psiquiatras diagnostican manualmente los trastornos mentales, incluida la depresión. Los métodos actuales consisten en entrevistas y cuestionarios como PHQ-2 (Kroenke et al., 2003), PHQ-8 (Kroenke et al., 2009) y PHQ-9 (L. Costantini et al., 2021), los cuales, están diseñados para diagnosticar este tipo de trastorno. Los resultados de estas encuestas se analizan por psiquiatras, que posteriormente realizarán una entrevista individual con el paciente. Durante la entrevista, los psiquiatras buscan marcadores relacionados con la depresión en el habla de los pacientes, por ejemplo, exhibición emocional, razonamientos e inconsistencias (Manoret et al., 2021).

Se sabe que el diagnóstico manual de la depresión es eficaz, pero no hay suficientes psiquiatras disponibles y el diagnóstico en persona suele llevar mucho tiempo. Además, la carga financiera en los trastornos de depresión mayor es relativamente alta (L. Costantini et al., 2021). Afortunadamente, se han desarrollado varias tecnologías médicas para abordar estos problemas.

Actualmente, ha habido una integración de la Inteligencia Artificial en varios campos del análisis médico (Tătaru et al., 2021), pero no se utilizan en la detección de trastornos psicológicos. La IA puede aprender y dominar las habilidades desarrolladas por los psiquiatras, por ejemplo, el análisis de patrones del habla. Por lo tanto, la IA se ha convertido en una alternativa prometedora a la detección manual de la depresión.

El PHQ-8 es un cuestionario de salud del paciente de ocho ítems comúnmente utilizado para medir la gravedad de la depresión (Kroenke et al., 2009). Cada pregunta dentro del cuestionario pregunta con qué frecuencia los encuestados experimentan ciertos síntomas que se encuentran comúnmente en pacientes deprimidos. La frecuencia de estos síntomas se correlaciona directamente con la puntuación otorgada a cada pregunta. Aquellos que no experimentan los síntomas en absoluto reciben una puntuación de 0 y aquellos que experimentan el síntoma casi a diario reciben una puntuación de 3, esto se detalla en la Tabla 5.

**Tabla 5**

*Puntuaciones de PHQ-8 y gravedad de la depresión*

| Puntuación PHQ-8 | Nivel de severidad |
|---|---|
| 0-4 | Depresión mínima |
| 5-9 | Depresión ligera |
| 10-14 | Depresión moderada |
| 15-19 | Depresión moderadamente severa |
| 20-24 | Depresión severa |

*Fuente:* (Kroenke et al., 2009)

El habla es un acto de expresión de ideas y emociones mediante la vocalización (Manoret et al., 2021). También es un componente indispensable para la comunicación entre individuos dentro de la sociedad humana. En cuanto a la comunicación, junto con el habla se ha utilizado otro elemento llamado “lenguaje”. El lenguaje es la forma de expresar el

pensamiento a través de un conjunto distinto de símbolos, dialectos o sonidos (habla). La comprensión del idioma se puede adquirir mediante un estudio exhaustivo de los patrones vocales y los alfabetos. Los seres humanos son capaces de identificar y expresar el habla y los idiomas; mientras que, las máquinas no tienen la capacidad de hacerlo (Manoret et al., 2021).

Para el Reconocimiento Automático de Voz (Automatic Speech Recognition: ASR), el sistema procesa los datos vocales (habla) en señales digitales adecuadas para el entrenamiento y análisis de la inteligencia artificial (Haton, 2003). Los hablantes tienen patrones de voz únicos debido a la variación de personalidades y estructuras corporales. En consecuencia, el ASR utiliza criterios como el tamaño del habla y los estilos de habla para clasificar las muestras de voz en grupos. Los espectrogramas y las técnicas de características cromáticas pueden mejorar potencialmente el sistema en la organización de las voces (Haton, 2003). Ambas técnicas extraen y presentan características relevantes para el sistema, lo que le permite realizar clasificaciones y evaluaciones más complejas (Bailey & Plumbley, 2021).

### 2.12. Test de Beck

El Inventario de Depresión de Beck (Beck Depression Inventory: BDI) es uno de los métodos de autoevaluación más válidos para medir la gravedad de la depresión. Su segunda edición (BDI-II) consta de 21 ítems que abordan síntomas cognitivos, afectivos, motivacionales y fisiológicos de la depresión, que pueden ser utilizados en adultos y adolescentes mayores de 13 años. El BDI-II es fácil de asignar, rentable y muy conveniente para la práctica clínica y la investigación. Muestra una validez de contenido adecuada en comparación con el Manual Diagnóstico y Estadístico de los Trastornos Mentales, Cuarta Edición, buena sensibilidad y especificidad moderada, confiabilidad test-retest satisfactoria y alta consistencia interna (Ciharova et al., 2020).

# CAPITULO 3

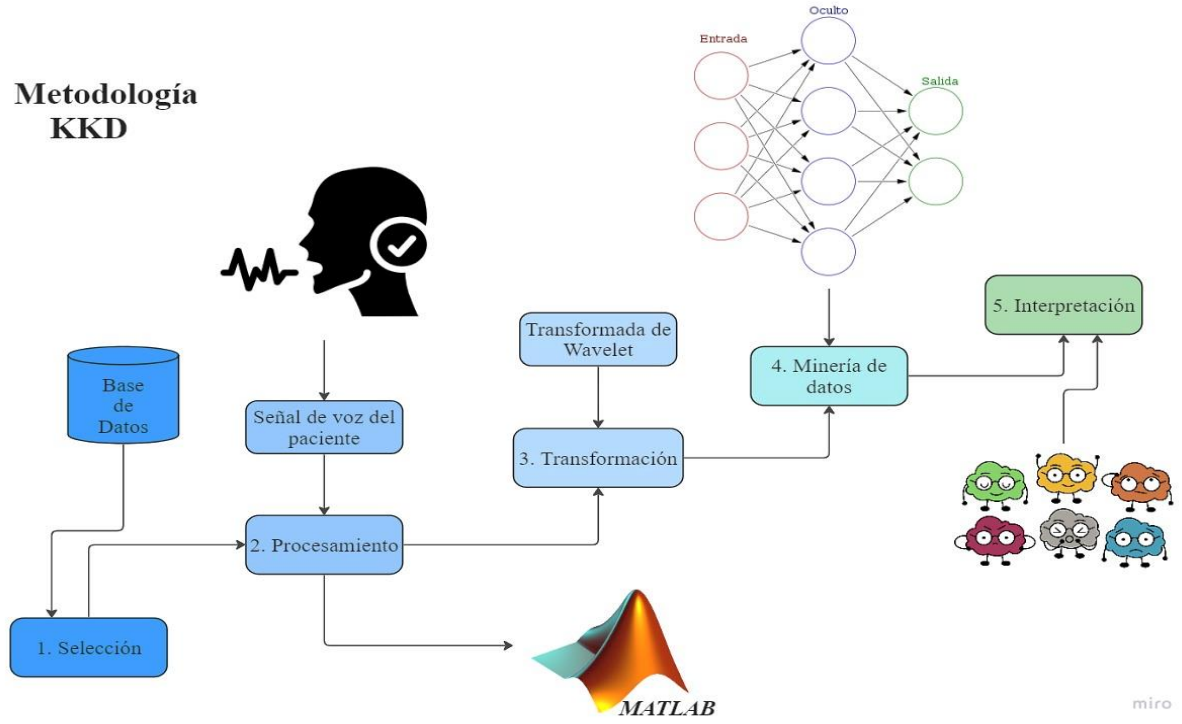
La mayor parte de las investigaciones se basan lo tradicional de extracción de características en el reconocimiento de emociones del habla, el principal problema de estas técnicas es el tiempo que conlleva el procesamiento de los audios.

En base a lo anterior, en esta investigación se propone el usó de un modelo de aprendizaje profundo de red neuronal con Memoria a Corto y Largo Plazo (Long Short Term Memory: LSTM), para el reconocimiento de emociones del habla. LSTM se caracteriza por ser rápido y preciso, por lo cual, esta técnica cumple con los requisitos de clasificación deseada.

Para la creación del sistema de reconocimiento de emociones del habla se contempla varios procesos: obtención de audios de emociones, preprocesamiento de señales, extracción de características, preparación de datos de entrenamiento y entrenamiento de la red neuronal. Todo este proceso se adapta a la metodología de descubrimiento de Conocimiento en Base de Datos (Knowledge Discovery in Databases: KDD) tal como se muestra en la Figura 18.

**Figura 18.**

*Proceso KDD generalizado*




*Fuente*: elaborado por el autor

### 3.1. Obtención de audios de emociones

Las tareas clásicas de minería de datos requieren los siguientes tipos de datos: un conjunto de datos de entrenamiento, utilizado para entrenar el algoritmo en los datos específicos del problema, y un conjunto de datos de prueba para evaluar la calidad del algoritmo. En contraste, la tarea en cuestión tiene como objetivo la clasificación binaria de emociones del habla, siendo las clases, ira, disgusto, miedo, felicidad, neutral y tristeza. Por lo tanto, durante la fase de recopilación de datos fue necesario generar al menos un audio por cada emoción humana. En contraste, debido a la carencia de arreglos de datos de audios en español se emplea un conjunto de audios mexicana denominada Base de Datos Mexicana del Discurso Emocional (Mexican Emotion Speech Database: MESD).

#### 3.1.1. MESD

La base de datos se la obtuvo del repositorio de Kaggle que es una comunidad científica orientada al aprendizaje automático, la licencia de la base de datos es de libre uso, sin embargo, se acredita al autor de dicha base que es Saurabh Shahane.

La Base de Datos del Discurso Emocional Mexicano proporciona enunciados de una sola palabra para las prosodias afectivas de ira, disgusto, miedo, felicidad, neutro y tristeza con conformación cultural mexicana. El MESD ha sido pronunciado por actores adultos y niños no profesionales: Se dispone de 3 voces femeninas, 2 masculinas y 6 infantiles (edad media femenina de = 23,33 ± 1,53, edad media masculina de = 24 ± 1,41 y edad media infantil de = 9,83 ± 1,17). Las palabras de los enunciados emocionales y neutros proceden de dos corpus: (corpus A) compuesto por sustantivos y adjetivos que se repiten en todas las prosodias emocionales y tipos de voz (femenina, masculina, infantil), y (corpus B) que consiste en palabras controladas por edad de adquisición, frecuencia de uso, familiaridad, concreción, valencia, excitación y clasificaciones discretas de dimensionalidad de la emoción.

Las grabaciones de audio se realizaron en un estudio profesional con los siguientes materiales: (1) un micrófono Sennheiser e835 con una respuesta en frecuencia plana (100 Hz a 10 kHz), (2) una interfaz de audio Focusrite Scarlett 2i4 conectada al micrófono con un cable de Línea de Retorno Externa (eXternal Line Return: XLR) y al ordenador, y (3) la estación de trabajo de audio digital (Rapid Environment for Audio Production, Engineering, and Recording: REAPER). Los archivos de audio se almacenaron como una secuencia de 24 bits con una frecuencia de muestreo de 48000Hz.

La cantidad de audios de emociones para entrenar a la red neuronal LSTM debe ser significativa, no obstante, el proceso de adquisición de audios de emociones ya sea mediante sesiones de voz pregrabadas o audios de chat que demuestren distintas emociones, conlleva mucho tiempo. Por lo cual, se debe buscar en bases de datos de emociones del habla previamente tomadas, para obtener el conjunto de audios necesarios y ahorrar tiempo, Es así como, se considera una base de datos que contiene 864 grabaciones de voz con seis prosodias diferentes: ira, disgusto, miedo, felicidad, neutral y tristeza. Además, se incluyen tres categorías de voz: mujer adulta, hombre adulto y niño, esto se observa en la Tabla 6.

**Tabla 6**.

*Prosodias de base de datos*

| | |
|---|---|
| **Base de datos** | MESD |
| **Interprete** | Hombres, Mujeres, Niños/as |
| **Tamaño** | 864 grabaciones |
| **Prosodias** | Ira, disgusto, miedo, felicidad, neutral, tristeza |
| **Licencia** | Gratuita |
| **Frecuencia de muestreo** | 48kHz |

*Fuente:* (MESD, 2022)

Una característica muy importante de los audios obtenidos es que cada uno posee resolución y calidad estandarizada de 24 bits con una frecuencia de muestreo de 48000Hz. Por consiguiente, los nuevos audios deben ser preprocesados de alguna manera, para que cumplan con los requisitos de la red neuronal. El preprocesamiento de audios es muy importante, debido a que, el rendimiento de la red neuronal y la tasa promedio de registro de detección depende en gran medida de los datos de entrada, que se emplean para el entrenamiento de la red. Por lo que, se procura que estos sean claros y precisos.

### 3.1.2. Señales de audio de cada emoción

Cada emoción tiene características específicas al momento de pronunciarlas, esto se puede ver reflejado realizando la gráfica de la señal de audio, para ello se realiza un análisis de los datos obtenidos en bruto, es decir, se analiza las gráficas provenientes de cada emoción para destacar las características importantes de cada una de ellas. Por consiguientes, se escoge el audio de “Arriba”, que se ha expresado en los seis estados de ánimo (ira, disgusto, miedo, felicidad, neutral y tristeza).

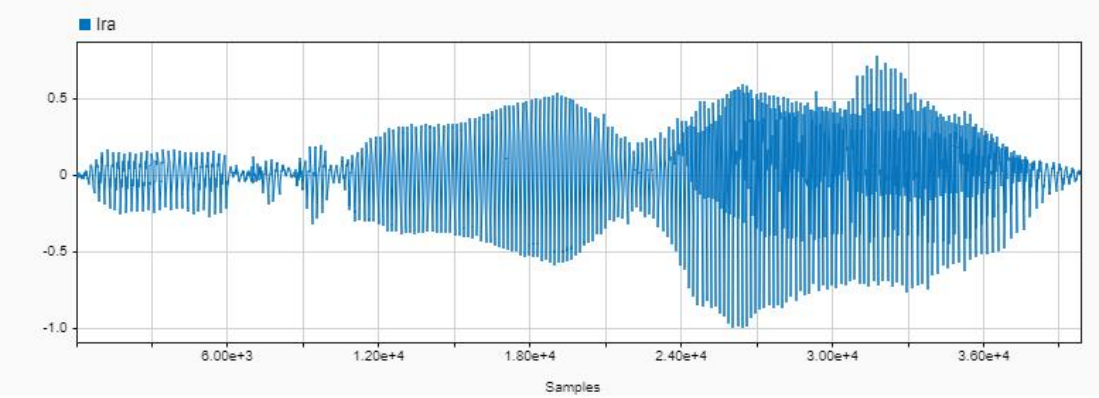
#### *3.1.2.1. Ira*

En la Figura 19 (a) se puede apreciar la señal que corresponde a la ira, en la primera parte de la figura se tiene la señal en función del tiempo, y la Figura 19 (b) el espectro de la señal, por consiguiente se observa el comportamiento de la voz mientras se tiene la emoción de la ira, la amplitud de la señal empieza con picos cortos y se segmenta en una parte de las muestras, luego los picos en amplitud van creciendo y al final se observa que la señal llega a picos muy altos tanto en el eje positivo como en el negativo, esto se mantiene constante hasta terminar la señal, esta señal se la compara con la forma en que se expresa la ira de forma vocal, ya que se inicia en un tono suave y al final se termina en un tono alto de voz produciendo la señal que se mira en el gráfico. La energía de la señal llega hasta los -20dB como se observa en el grafico del espectro.

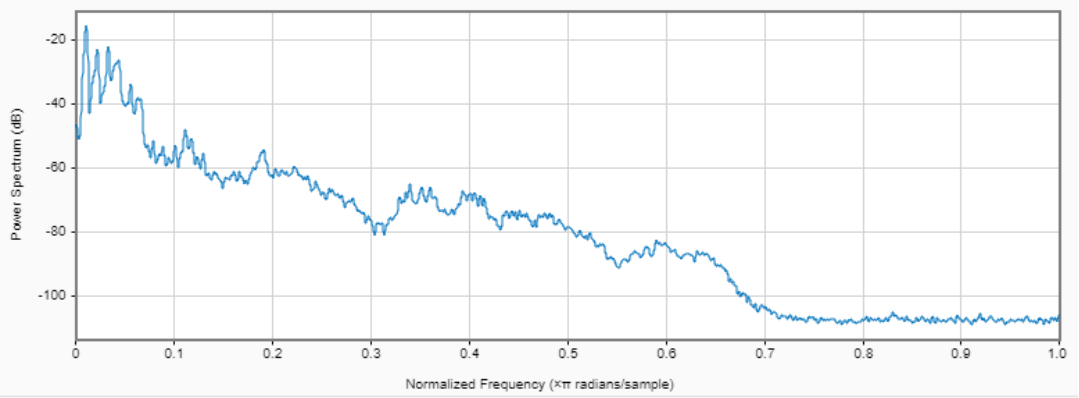
**Figura 19.**

*Gráfica en función del tiempo y espectro de la señal de Ira*

*a) Gráfica en función del tiempo de la señal de la ira.*




*b) Gráfica del espectro de la señal de ira*




*Fuente:* Elaborado por el autor

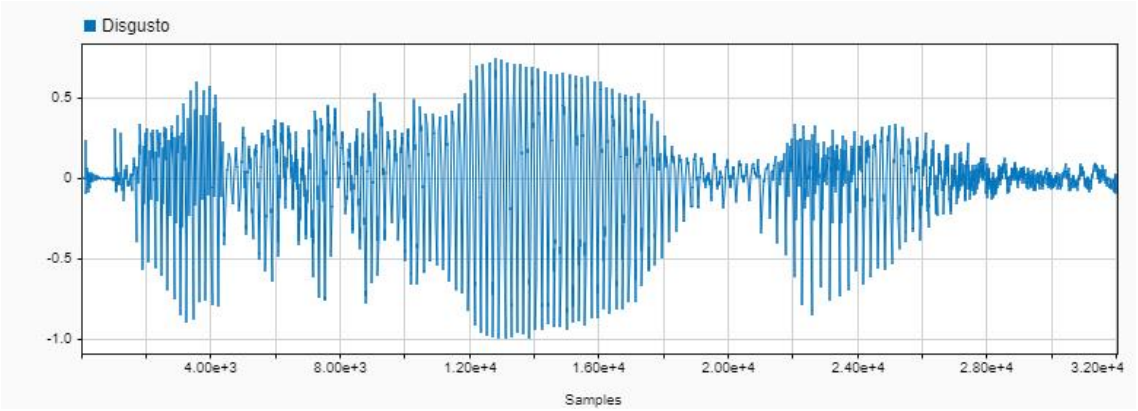
### *3.1.2.2. Disgusto*

El disgusto es una emoción negativa que con lleva altos niveles de amplitud en la señal, en la Figura 20 (a) se observa las gráficas para esta emoción en la cual se tiene que empieza con picos altos en amplitud, y conforme las muestras pasan a través del tiempo van segmentándose en tres ocasiones hasta obtener una señal constante en amplitud alta y finaliza con una amplitud constante baja con ciertos niveles inconsistentes, la señal se relaciona con la forma de expresar la emoción ya que la palabra se la pronuncia con poca energía. En el espectro de la señal se tiene que la energía se mantiene constante en la mayor parte de la señal como se observa en la Figura 20 (b).

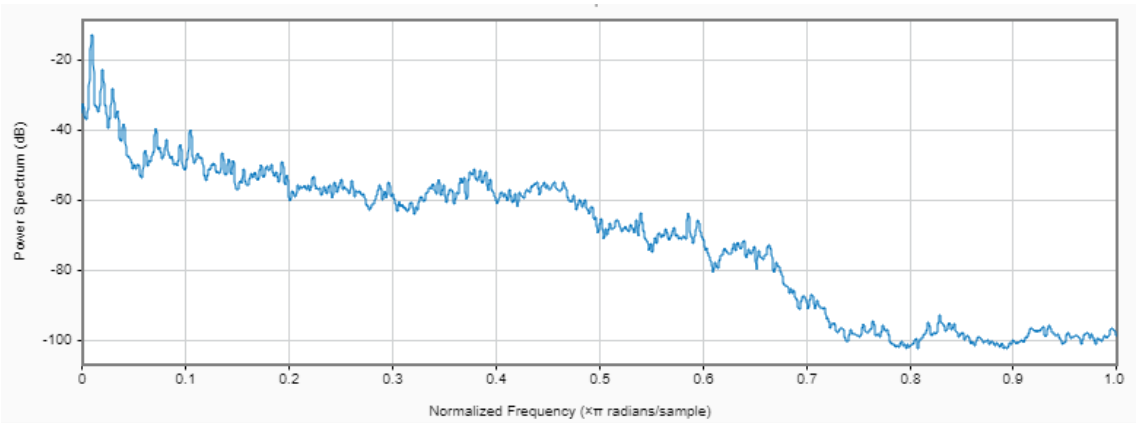
**Figura 20.**

*Gráfica en función del tiempo y espectro de la señal de Disgusto*

*a) Gráfica en función del tiempo de la señal de disgusto*




*b) Gráfica del espectro de la señal de disgusto*




*Fuente:* Elaborado por el autor

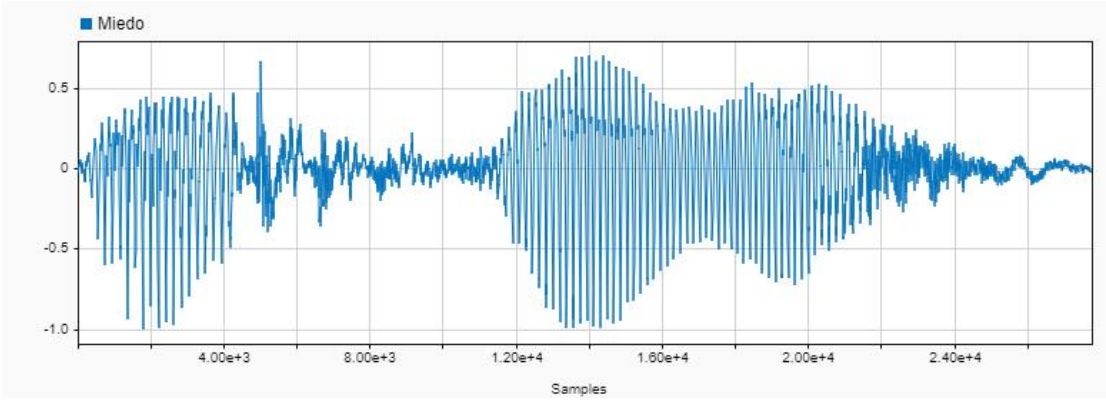
### *3.1.2.3. Miedo*

El miedo es una emoción que está caracterizada por la experimentación de una sensación desagradable o que se note la precepción de un peligro ya sea real o imaginario, en la Figura 21 (a) se aprecia la señal del miedo, la característica principal para esta emoción es que se tiene largos periodos de amplitud baja, tendiendo a cero, y solo dos fragmentos de señal en donde la amplitud es alta y constante, esto lo relacionamos con la forma de expresar el miedo en cada persona, las silabas se entre cortan y existen espacios de silencios que conforman esta emoción, en contraste se tiene la gráfica del espectro de la señal, se puede apreciar que el nivel de energía disminuye significativamente y no se tiene una señal que se mantenga constante Figura 21 (b).

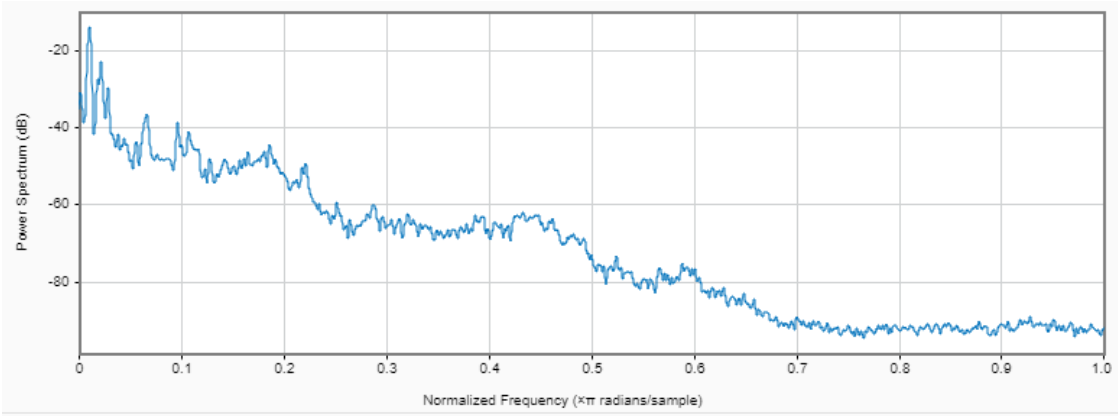
**Figura 21.**

*Gráfica en función del tiempo y espectro de la señal de Miedo*

*a) Gráfica en función del tiempo de la señal de miedo*




*b) Gráfica del espectro de la señal de miedo*




*Fuente:* Elaborado por el autor

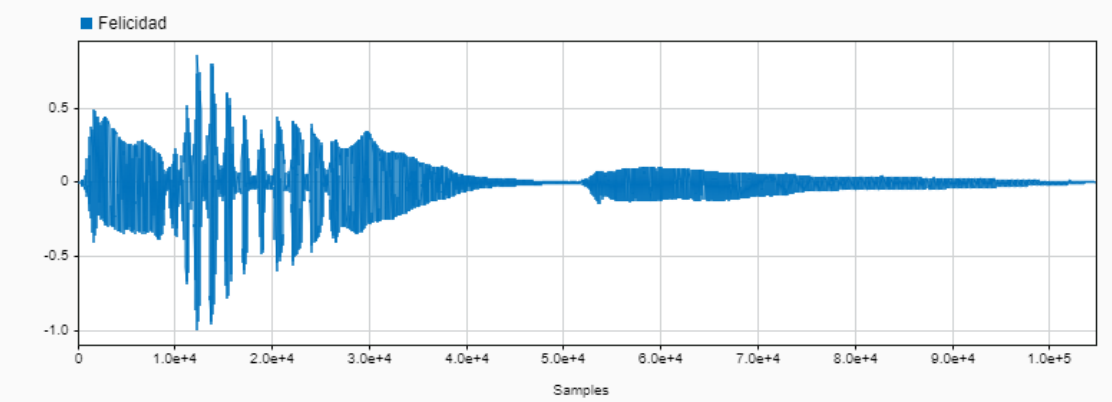
### *3.1.2.4. Felicidad*

La felicidad es una de las emociones en donde se puede apreciar mayores características, como muchos picos en amplitud seguidos de nodos en la señal, al empezar la señal se obtiene una amplitud estable y luego se aprecian estos picos y nodos en amplitud finalizando con una constante en amplitud, mirar Figura 22 (a), así podemos comparar con el habla al momento de expresar esta emoción, ya que se lo hace con una velocidad de locución rápida y altos niveles de potencia, esto se ve reflejado en la Figura 22 (b) del espectro de la señal que mantiene una potencia alta y al final decrece como todas las demás emociones.

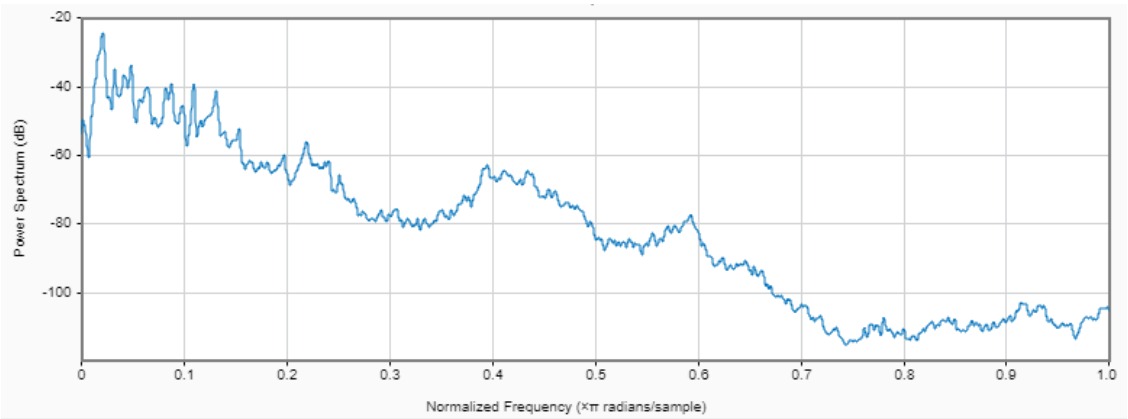
**Figura 22.**

*Gráfica en función del tiempo y espectro de la señal de Felicidad*

*a) Gráfica en función del tiempo de la señal de felicidad*




*b) Gráfica del espectro de la señal de felicidad*




*Fuente:* Elaborado por el autor

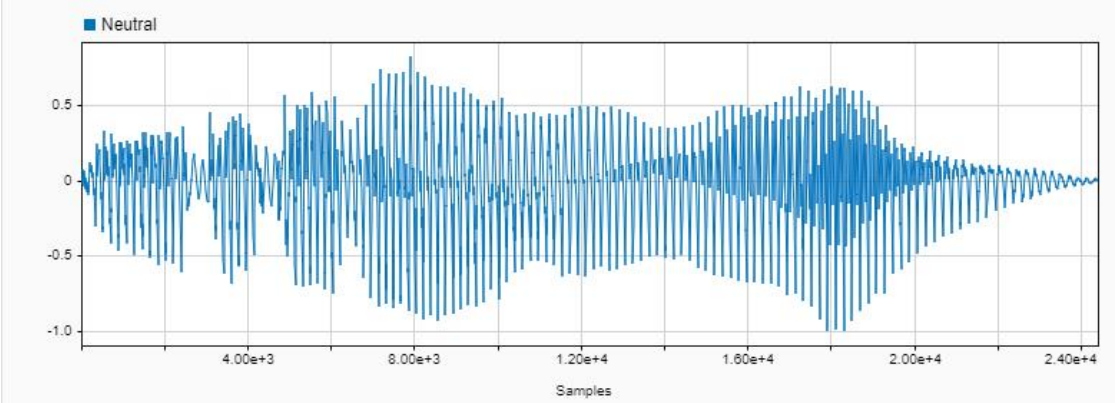
### *3.1.2.5. Neutral*

EL estado neutral se refiere a cuando se mantiene un tono de voz constante, esta es la señal que más predomina al momento de establecer una conversación. Se observa una constante en amplitud a lo largo de la onda, con excepciones que se hacen visible en el inicio de la onda, mirar Figura 23 (a). Es también visible en el espectro de la señal en donde se tiene variantes en la potencia a lo largo de la misma, el gráfico tiene en común la forma de expresarse a lo largo de una conversación en donde se tiene un tono de voz normal y neutral en el cual no se expresa ninguna emoción, tal como se aprecia en la Figura 23 (b).

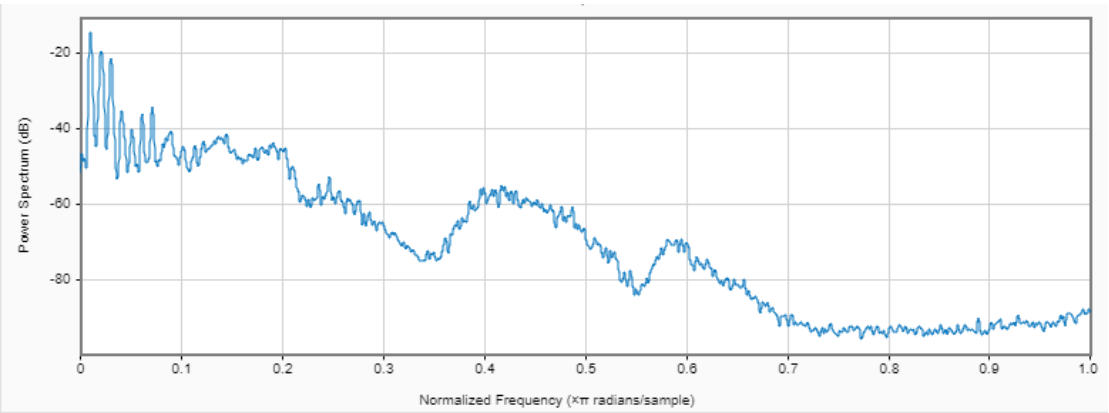
**Figura 23.**

*Gráfica en función del tiempo y espectro de la señal Neutral*

a) *Gráfica en función del tiempo de la señal neutral*




b) Gráfica del espectro de la señal neutral




*Fuente:* Elaborado por el autor

### *3.1.2.4. Tristeza*

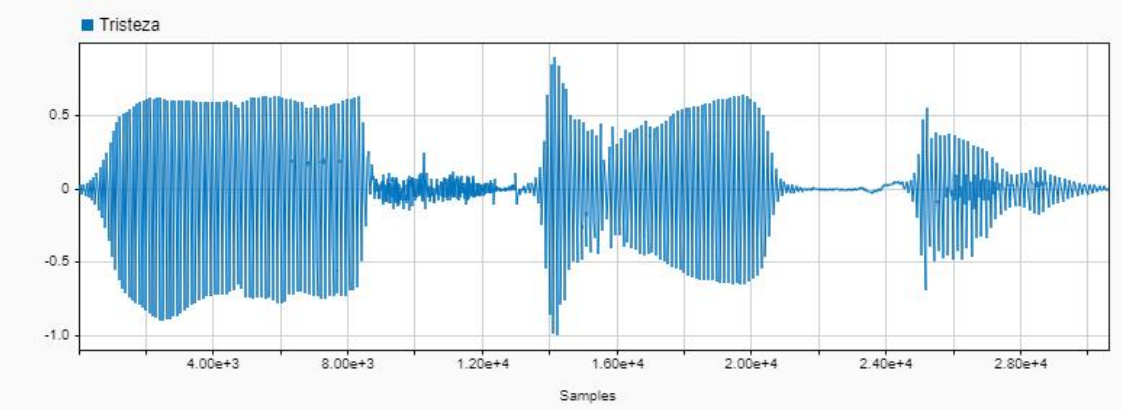
Esta emoción es en la que más se puede apreciar características identificativas para asociar con la emoción, el comportamiento de la señal varía mucho, en primer lugar se tiene un flujo constante de muestras que tienen una amplitud alta y luego se aprecia que se tiene una amplitud baja, esto se repite por dos periodos finalizando con una constante de amplitud media, esa señal se relaciona con la emoción ya que al momento de expresar tristeza se mantiene cortos periodos de tiempo en silencio debido al nerviosismo y timidez que genera esta emoción Figura 24 (a).

En el espectro se aprecia un gran decremento del nivel de potencia iniciando cerca de los -19dB y acelerando este decremento cerca de los -60dB, observar Figura 24 (b).

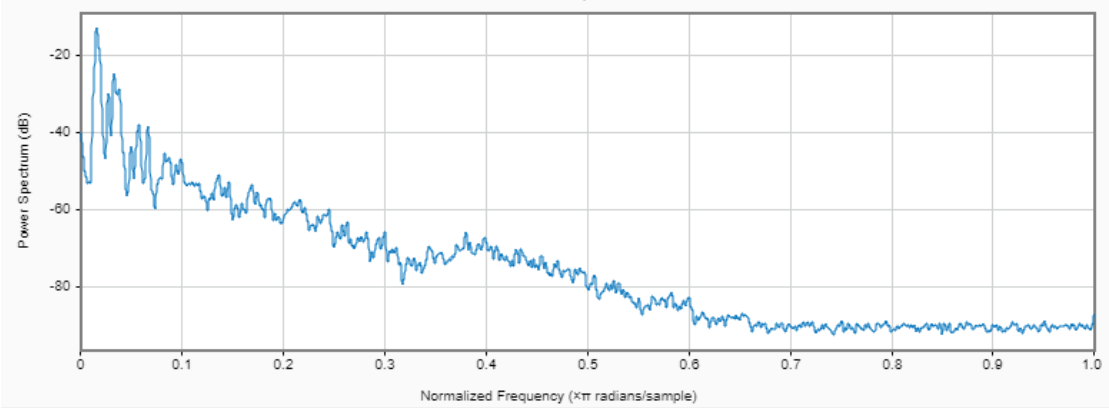
**Figura 24.**

*Gráfica en función del tiempo y espectro de la señal de Tristeza*

*a) Gráfica en función del tiempo de la señal de tristeza*




*b) Gráfica del espectro de la señal de tristeza*




*Fuente:* Elaborado por el autor

### 3.2. Preprocesamiento de señales

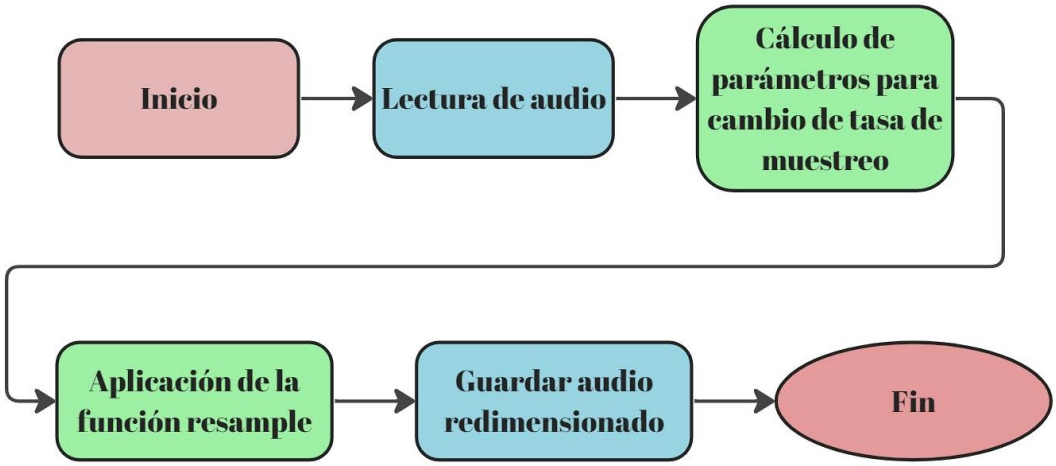
El preprocesamiento se refiere a algún procedimiento previo que se realiza con los audios, estos pueden ser, redimensionamiento, cambio de tasa de muestreo, aumento de calidad y extracción de características.

Algo muy importante que se debe mencionar es que el ancho de banda de los audios de la base de datos es de 48000Hz, no obstante, posterior al proceso de redimensionamiento, los audios poseen un ancho de banda de 16000 Hz. Por consiguiente, la base de datos consta de audios que poseen este último atributo.

Para el caso de los audios obtenidos en el apartado anterior se realiza una revisión de la tasa de muestreo que posee cada uno de los elementos. Esto se realiza para garantizar que todos los audios posean parámetros estandarizados. Para tal efecto, se emplea la función de volver a muestrear que posee MATLAB. Esta función cambia la tasa de muestreo de un audio a una tasa de muestreo deseada. Esto se lleva a cabo mediante el proceso que se presenta en la Figura 25.

**Figura 25.**

*Proceso para cambio de tasa de muestreo de audios*



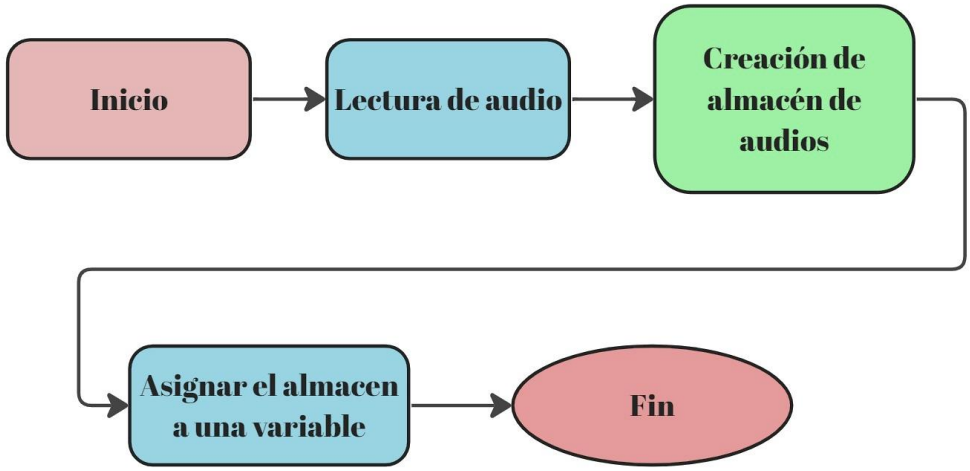

*Fuente:* Elaborado por el autor

En primer lugar, se lee el audio que se desea procesar, en segundo lugar, se aproximación de fracciones racionales mediante la función rat, luego se cambia la tasa de muestreo de la actual a la deseada aplicando la función de volver a muestrear y, por último, el audio resultante se reproduce y se guarda en una dirección deseada. Este proceso se repite para cada audio que se encuentre en la carpeta.

Posterior al cambio de tasa de muestreo se crea un almacén de datos con las direcciones y nombres de cada uno de los audios, esto se realiza para evitar la generación de variables extensas, en cuanto al peso en MB se refiere. Para lo cual, se emplea el proceso que se presenta en la Figura 26.

**Figura 26.**

*Proceso para crear el almacén de datos*




*Fuente:* Elaborado por el autor

### 3.3. Extracción de características

Este proceso consiste en identificar y extraer de forma numérica las características únicas de los audios de emociones. El proceso de extracción de características destacadas se realiza mediante el uso de la transformada de wavelet en Matlab. Para ello, se debe seguir una secuencia lógica de cinco pasos, entre los que se encuentran: lectura de audio, calcular la descomposición wavelet multinivel, cuantificar la entropía de la señal y asociar las características con la emoción correspondiente.

En contraste, la lectura de los audios se realiza mediante la función audioread, para seguir con la extracción de características mediate la transformada de wavelet.

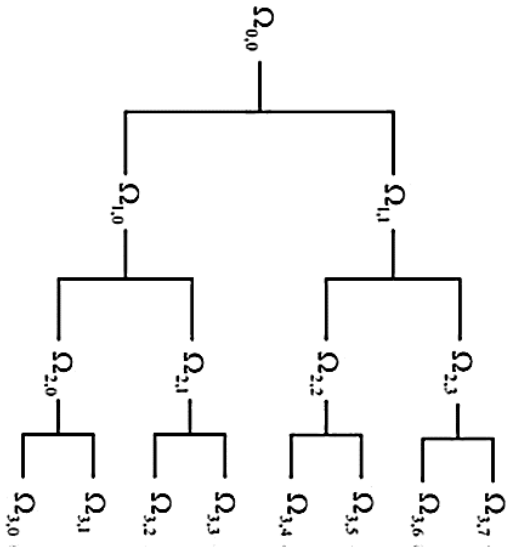
En este sentido, la transformada de Wavelet Packet puede considerarse como un árbol de subespacios, con $\Omega_{0,0}$ representando el espacio de señal original, es decir, el nodo raíz del árbol, mirar Figura 27.

La idea es descomponer la señal en una serie de subespacios utilizando las funciones wavelet elegidas (Symmlet y daubechies). Dicha descomposición generalmente se logra traduciendo (empujando hacia la derecha o hacia la izquierda) y escalando (amplificando o

atenuando) la función wavelet seleccionada, de esta forma, se proyecta la señal en el subespacio representado por la escala y traslación específicas.

**Figura 27.**

*Descomposiciones de paquetes wavelet de $\Omega_{0,0}$ en subespacios estructurados en el árbol*

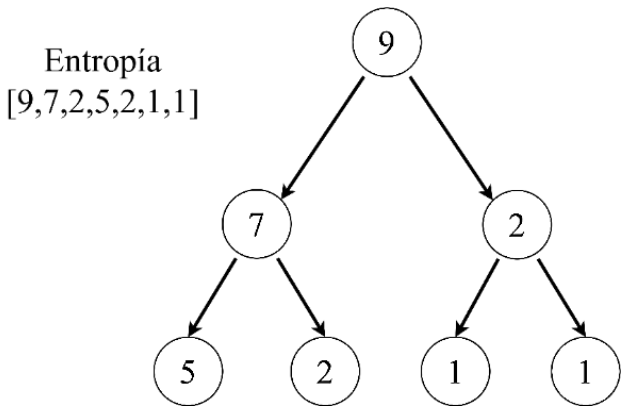



*Nota.* Adaptado de *(*Meng *et al., 2021)*

En el problema de reconocimiento de emociones se sabe que la energía de la señal del audio es una de las características más importantes. Por lo que, una solución simple para extraer las características es usar la energía de la señal que tiene en cada subespacio o nodo. Una vez que tenga el valor de la característica para cada nodo, se debe colocarlos todos uno al lado del otro y ese es su vector de características que se detalla en la Figura 28.

**Figura 28.**

*Ejemplo de extracción de características para tres niveles de descomposición*



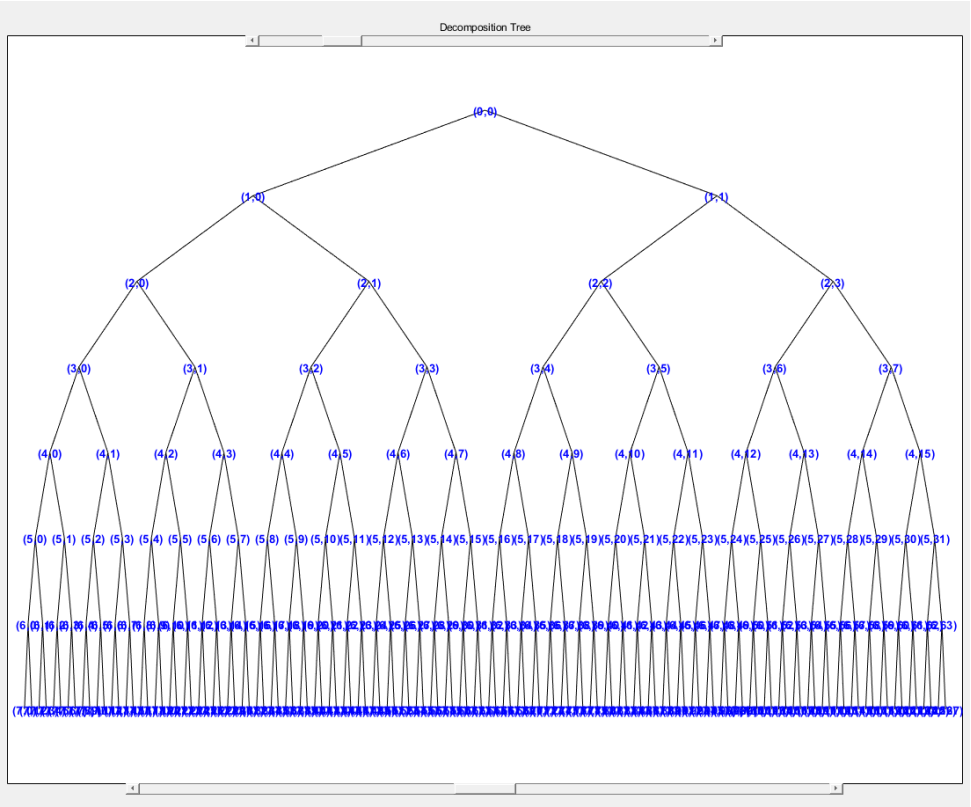

*Nota.* Adaptado de *(*Meng *et al., 2021)*

### 3.3.1. Transformada de Wavelet para cada emoción

En contraste, para ejemplificar el proceso descrito con anterioridad se procede a efectuar el cálculo de la transformada wavelet para una emoción de cada tipo considerado. Para la señal de ira se emplea una descomposición wavelet de la familia daubechies 4 con 7 niveles de descomposición. Por consiguiente, para siete niveles de composición se dispone de 255 características $(2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 + 2^6 + 2^7 = 255)$ tal como se observa en la Figura 29. En este sentido se presentan los dos primeros elementos del último nivel de descomposición, los cuales corresponden a los coeficientes de aproximación que se mira en la Figura 30 y detalle en la Figura 31. De manera que, con lo que respecta al coeficiente de aproximación se observa una señal mucho más clara a comparación de la señal de audio original, donde de manera pura, la amplitud se comporta de manera cambiante. Por otro lado, el coeficiente de detalle muestra una predominancia de la señal en las pausas del habla.

**Figura 29.**

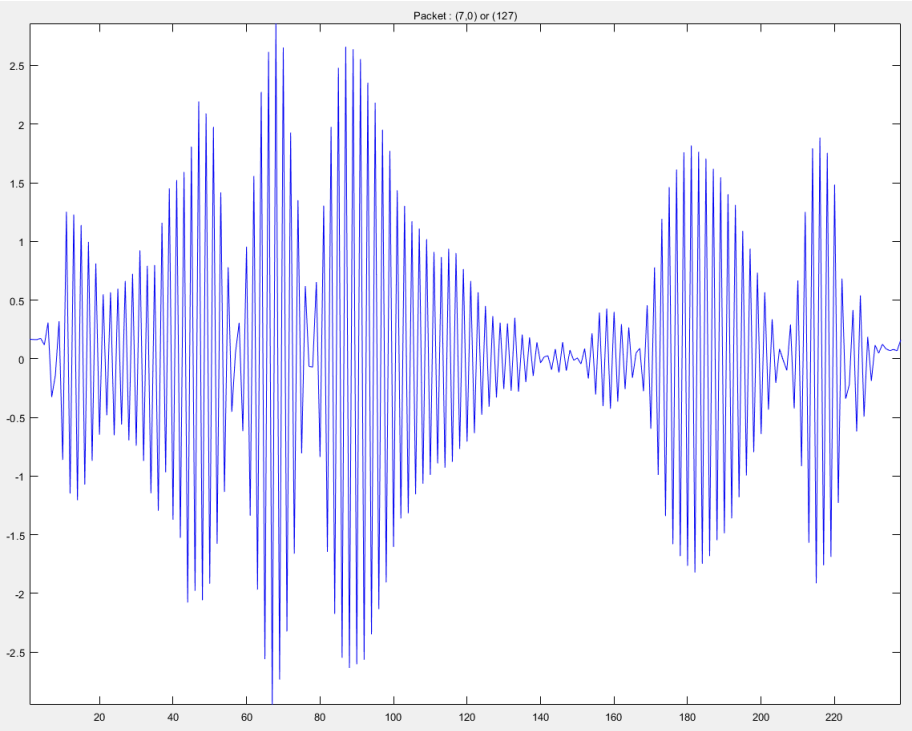
*Árbol de descomposición wavelet de 7 niveles para la emoción de la ira*




*Fuente:* Elaborado por el autor

**Figura 30.**

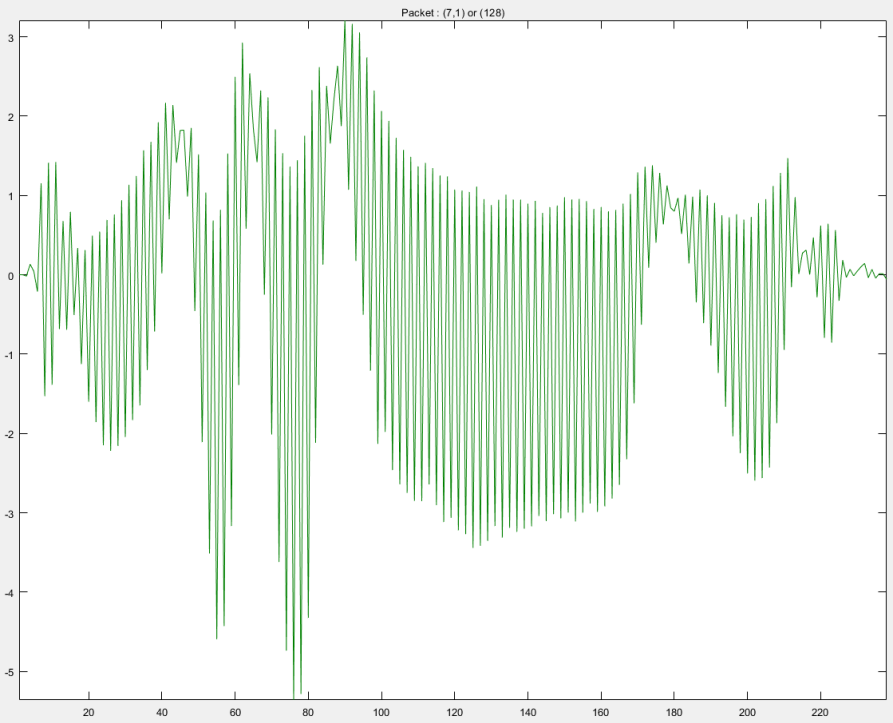
*Coeficiente de aproximación, nivel 7 asociado a la emoción de la ira.*




*Fuente:* Elaborado por el autor

**Figura 31.**

*Coeficiente de detalle, nivel 7 asociado a la emoción de la ira.*

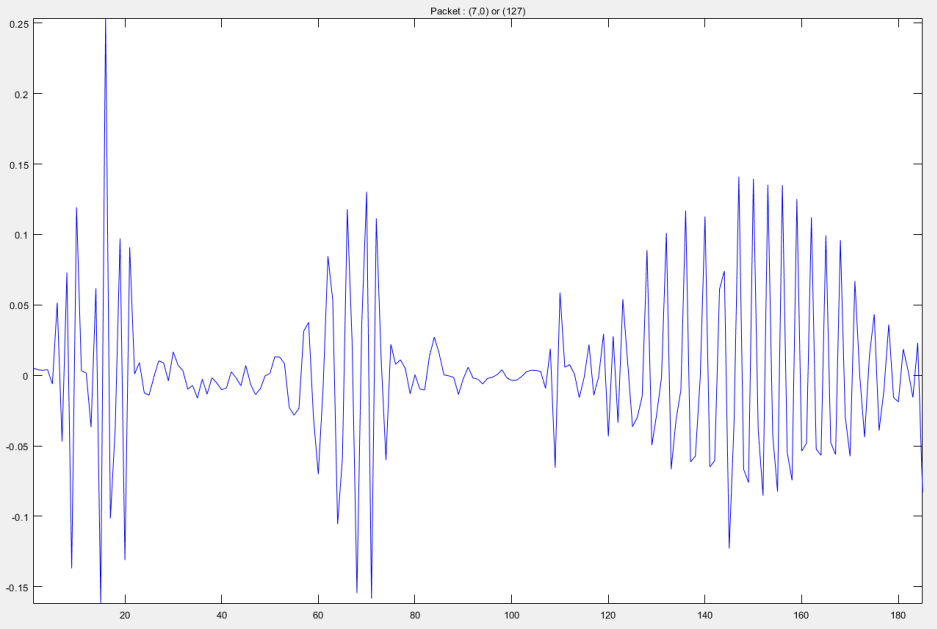



*Fuente:* Elaborado por el autor

Para la señal de disgusto se emplea una descomposición wavelet de la familia daubechies 4 con 7 niveles de descomposición. Por consiguiente, para siete niveles de composición se dispone de 255 características ($2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 + 2^6 + 2^7 = 255$). En este sentido se presentan los dos primeros elementos del último nivel de descomposición, los cuales corresponden a los coeficientes de aproximación que se observa en la Figura 32 y detalle en la Figura 33. En consecuencia, con lo que respecta al coeficiente de aproximación se observa una señal mucho más clara a comparación de la señal de audio original, donde se existen varias pausas. Por otro lado, el coeficiente de detalle muestra una predominancia una sola pausa en la onda, asimismo, la amplitud inicial crece de forma importante, para luego reducirse hasta la pausa.

**Figura 32.**

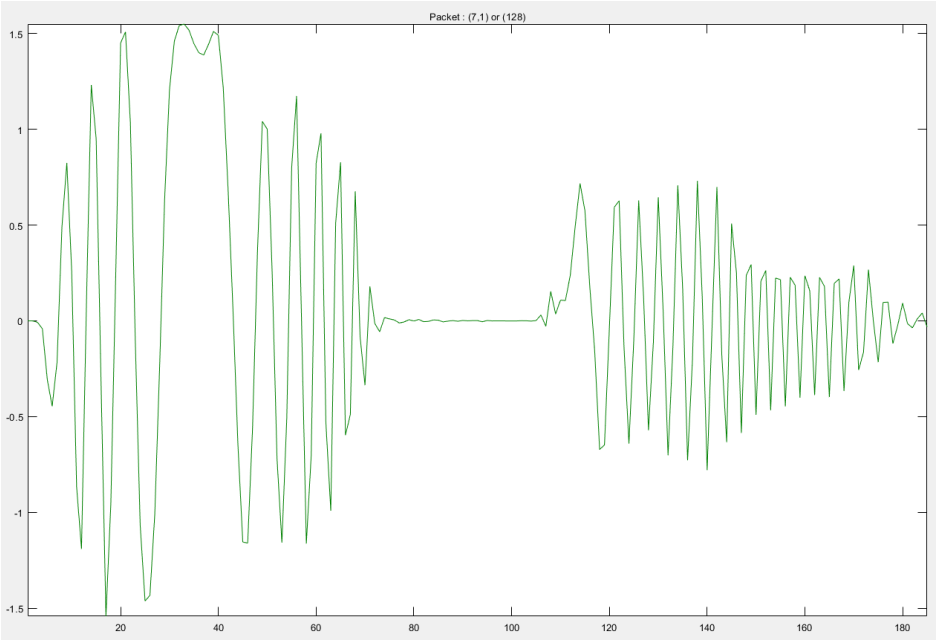
*Coeficiente de aproximación, nivel 7 asociado a la emoción de disgusto*




*Fuente:* Elaborado por el autor

**Figura 33.**

*Coeficiente de detalle, nivel 7 asociado a la emoción de disgusto*



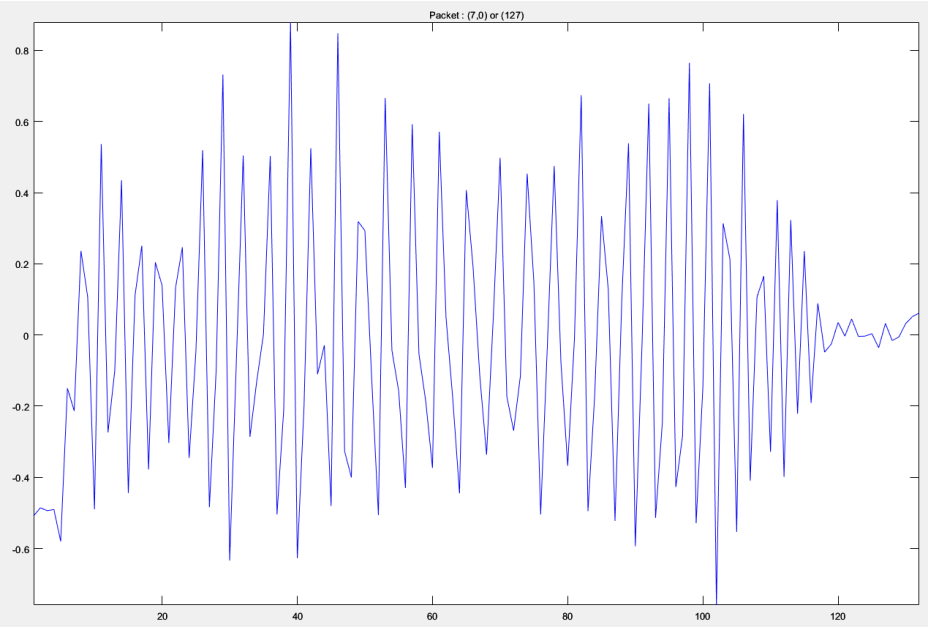

*Fuente:* Elaborado por el autor

Para la señal de miedo se emplea una descomposición wavelet de la familia daubechies 4 con 7 niveles de descomposición. Por consiguiente, para siete niveles de composición se dispone de 255 características ($2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 + 2^6 + 2^7 = 255$). En este sentido se presentan los dos primeros elementos del último nivel de descomposición, los cuales corresponden a los coeficientes de aproximación que se aprecia en la Figura 34 y detalle en la Figura 35.

En consecuencia, con lo que respecta al coeficiente de aproximación se observa una señal mucho más clara a comparación de la señal de audio original, donde la onda es constante sin presentar pausas. Por otro lado, el coeficiente de detalle muestra una onda con una amplitud creciente al inicio y decreciente al final.

**Figura 34.**

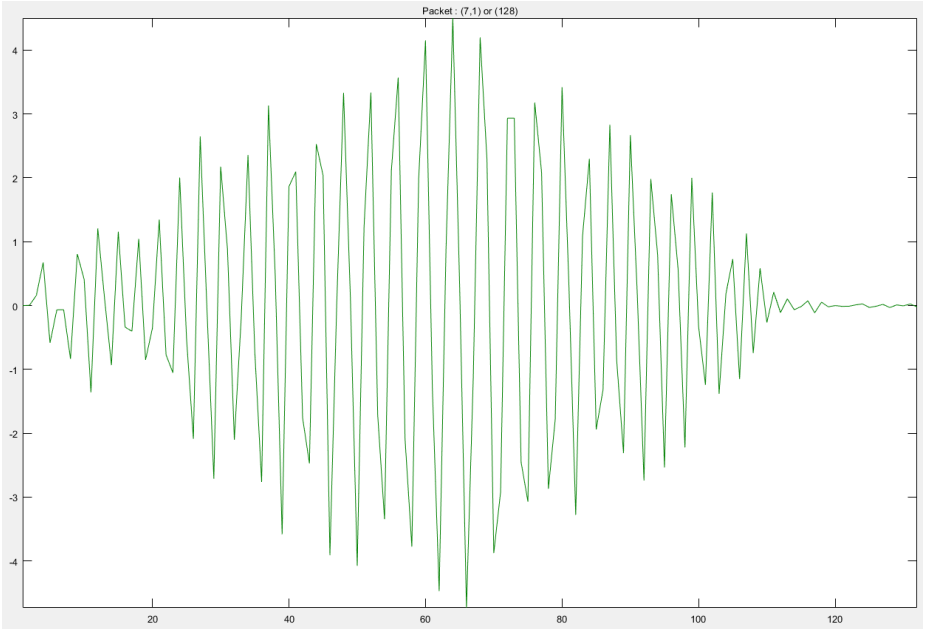
*Coeficiente de aproximación, nivel 7 asociado a la emoción de miedo*




*Fuente:* Elaborado por el autor

**Figura 35.**

*Coeficiente de detalle, nivel 7 asociado a la emoción de miedo*



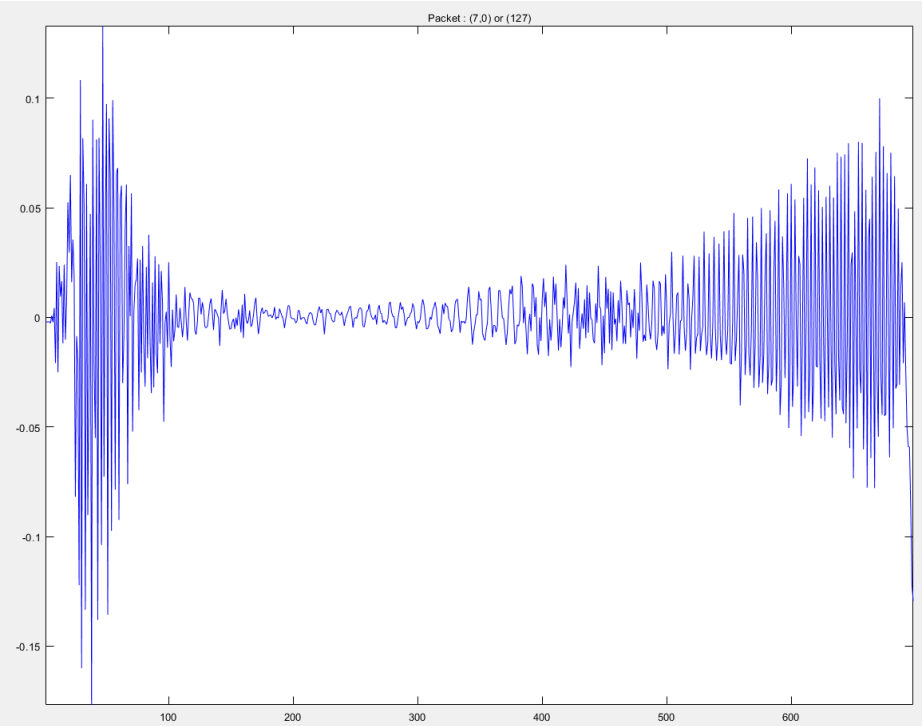

*Fuente:* Elaborado por el autor

Para la señal de felicidad se emplea una descomposición wavelet de la familia daubechies 4 con 7 niveles de descomposición. Por consiguiente, para siete niveles de composición se dispone de 255 características ($2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 + 2^6 + 2^7 = 255$). En este sentido se presentan los dos primeros elementos del último nivel de descomposición, los cuales corresponden a los coeficientes de aproximación que se observa en la Figura 36 y detalle en la Figura 37.

En consecuencia, con lo que respecta al coeficiente de aproximación se observa una señal mucho más clara a comparación de la señal de audio original, además, en primera instancia la onda posee una amplitud alta. Posteriormente, reduce la amplitud de forma considerable para volver a elevarse al final del tiempo. Por otro lado, el coeficiente de detalle muestra una onda con una amplitud constante, aunque la frecuencia varia.

**Figura 36.**

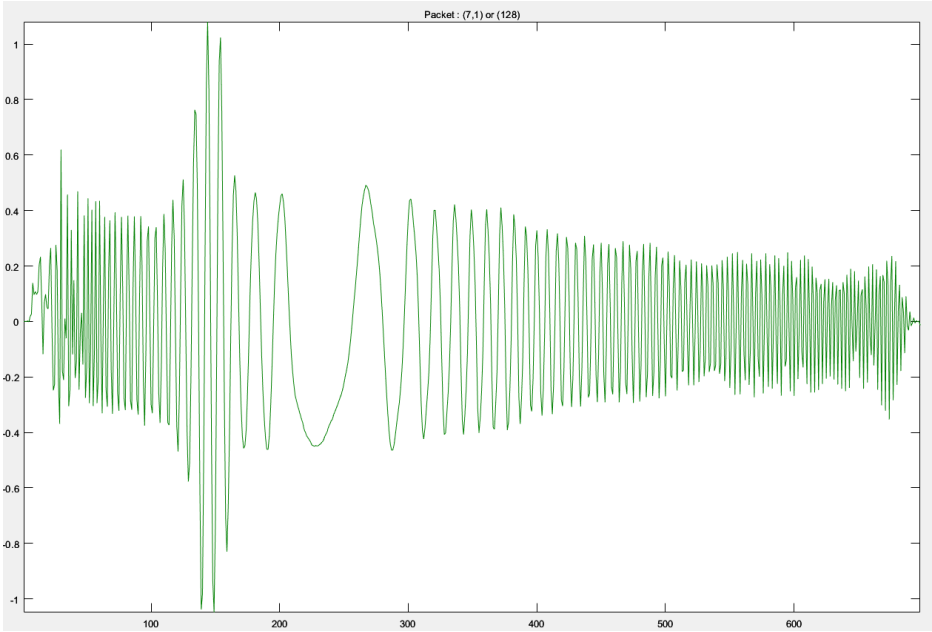
*Coeficiente de aproximación, nivel 7 asociado a la emoción de felicidad*




*Fuente:* Elaborado por el autor

**Figura 37.**

*Coeficiente de detalle, nivel 7 asociado a la emoción de felicidad*



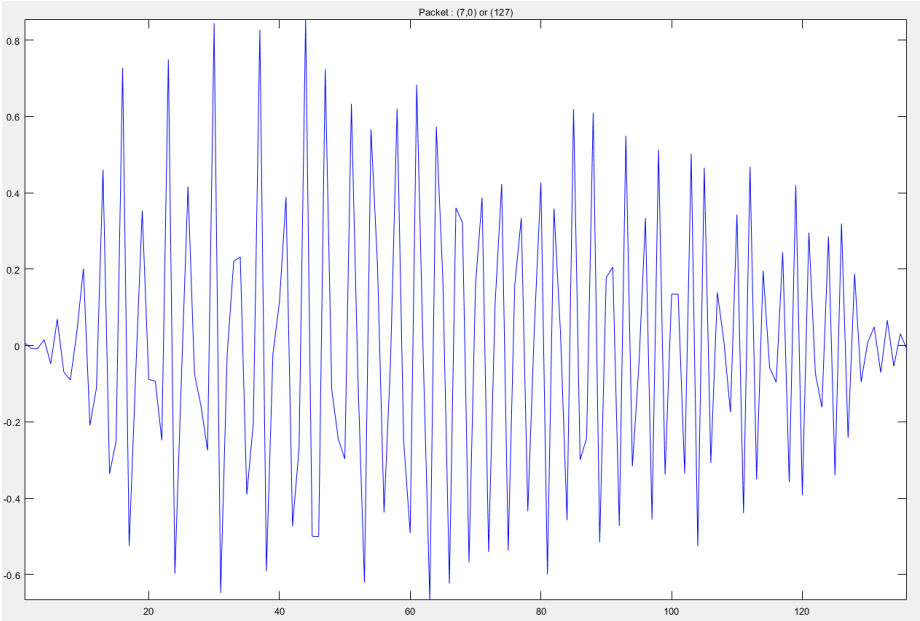

*Fuente:* Elaborado por el autor

Para la señal de felicidad se emplea una descomposición wavelet de la familia daubechies 4 con 7 niveles de descomposición. Por consiguiente, para siete niveles de composición se dispone de 255 características ($2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 + 2^6 + 2^7 = 255$). En este sentido se presentan los dos primeros elementos del último nivel de descomposición, los cuales corresponden a los coeficientes de aproximación se aprecian en la Figura 38 y detalle en la Figura 39.

En consecuencia, con lo que respecta al coeficiente de aproximación se observa una señal mucho más clara a comparación de la señal de audio original, además, la amplitud se mantiene uniforme de inicio a fin. Por otro lado, el coeficiente de detalle muestra una onda con una amplitud que se reduce de forma paulatina al final.

**Figura 38.**

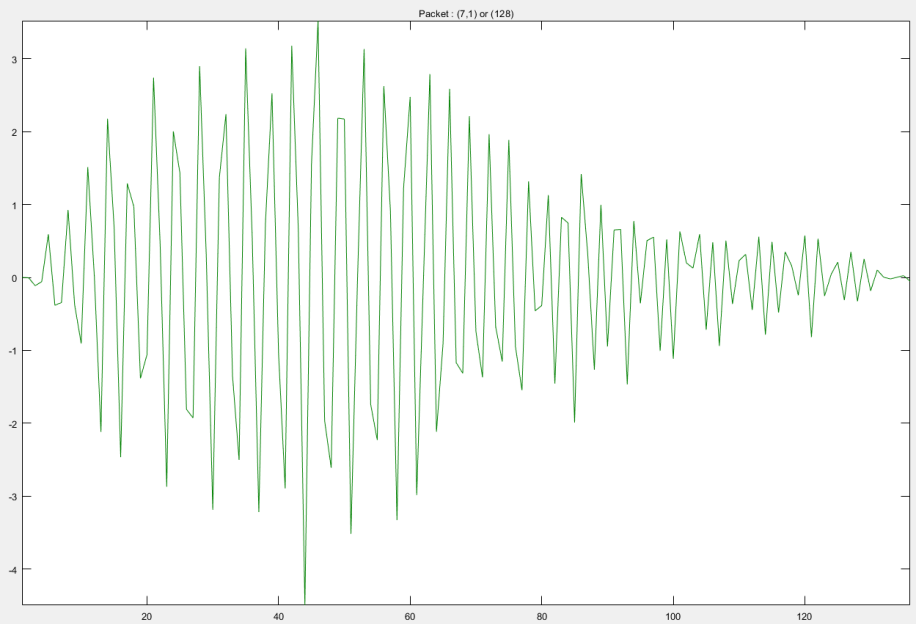
*Coeficiente de aproximación, nivel 7 asociado a la emoción de estado neutral*




*Fuente:* Elaborado por el autor

**Figura 39.**

*Coeficiente de detalle, nivel 7 asociado a la emoción de estado neutral*



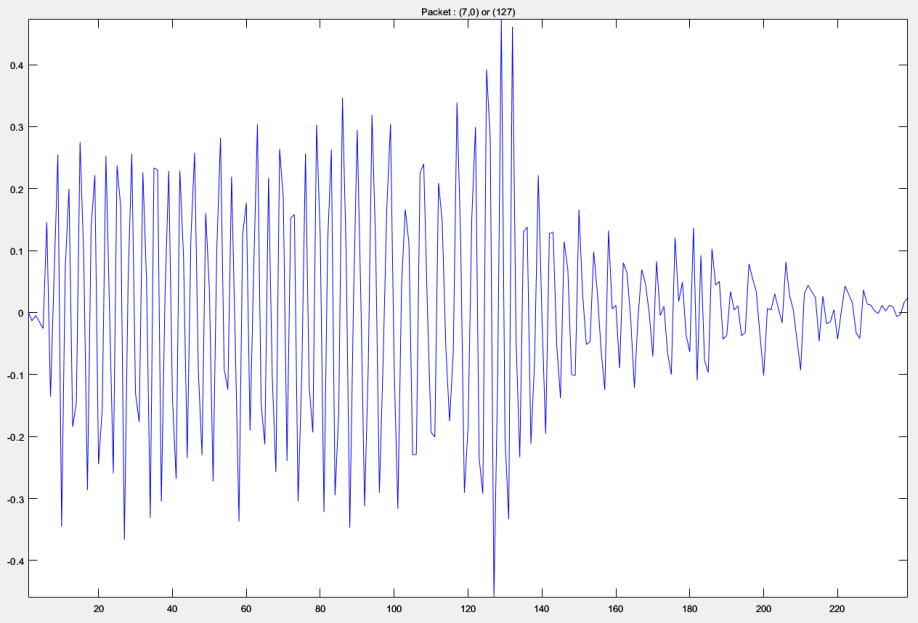

*Fuente:* Elaborado por el autor

Para la señal de tristeza se emplea una descomposición wavelet de la familia daubechies 4 con 7 niveles de descomposición. Por consiguiente, para siete niveles de composición se dispone de 255 características $(2^0 + 2^1 + 2^2 + 2^3 + 2^4 + 2^5 + 2^6 + 2^7 = 255)$. En este sentido se presentan los dos primeros elementos del último nivel de descomposición, los cuales corresponden a los coeficientes de aproximación se observan en la Figura 40 y detalle en a Figura 41.

En consecuencia, con lo que respecta al coeficiente de aproximación se observa una señal mucho más clara a comparación de la señal de audio original, además, la amplitud se mantiene uniforme al inicio. No obstante, al finalizar el ciclo la amplitud disminuye de forma paulatina. Por otro lado, el coeficiente de detalle muestra una onda con una amplitud constante, excepto en la mitad del audio.

**Figura 40.**

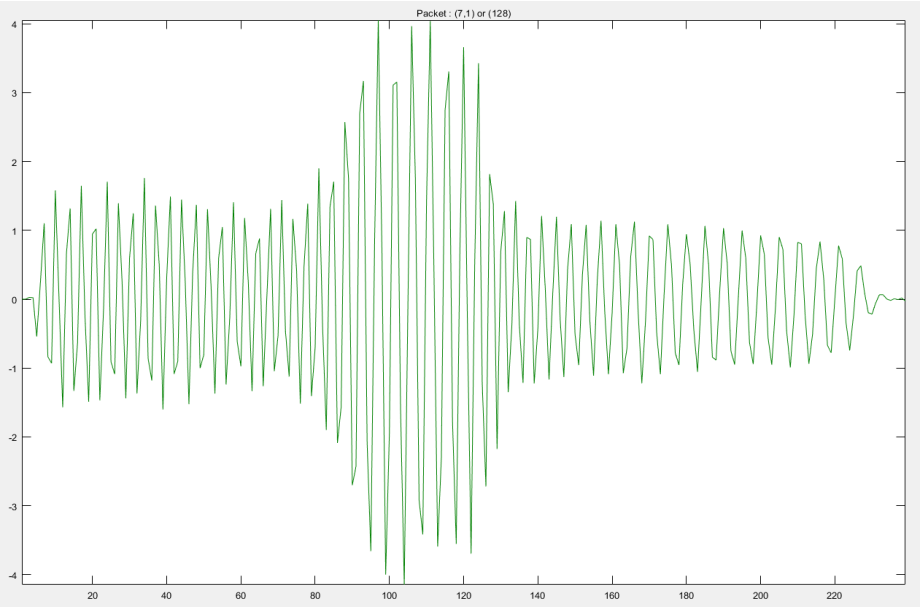
*Coeficiente de aproximación, nivel 7 asociado a la emoción de tristeza*




*Fuente:* Elaborado por el autor

**Figura 41.**

*Coeficiente de detalle, nivel 7 asociado a la emoción de tristeza*




*Fuente:* Elaborado por el autor

Una vez que obtiene la extracción de características de cada uno de los audios que se encuentran en el data set se procede a realizar el cálculo de la energía de cada señal resultante de la descomposición Wavelet multinivel, para este caso se usó un nivel 7 por ende se tiene 255 energías provenientes de cada señal, por consecuente una emoción será representada por 255 características, esto se muestra en la Figura 42 en donde está el data set con las características de todas las señales de audios de emociones que se obtiene del MESD.

**Figura 42.**

*Conjunto de datos de entrenamiento*

car

868x255 double

| | 243 | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 857 | 3529e+03 | 4.2353e+03 | 4.3810e+03 | 4.3746e+03 | 4.4803e+03 | 3.8652e+03 | 4.0442e+03 | 4.0373e+03 | 4.1688e+03 | 4.2788e+03 | 4.3410e+03 | 4.3242e+03 | 4.4378e+03 |
| 858 | 2943e+03 | 4.2348e+03 | 4.3476e+03 | 4.3128e+03 | 4.4273e+03 | 3.8950e+03 | 4.0594e+03 | 4.0313e+03 | 4.1566e+03 | 4.3065e+03 | 4.3407e+03 | 4.2967e+03 | 4.4125e+03 |
| 859 | 7978e+03 | 3.6746e+03 | 3.7781e+03 | 3.8207e+03 | 3.9006e+03 | 3.3853e+03 | 3.5496e+03 | 3.5450e+03 | 3.6413e+03 | 3.7235e+03 | 3.7750e+03 | 3.8029e+03 | 3.8714e+03 |
| 860 | 7189e+03 | 3.7097e+03 | 3.7518e+03 | 3.7907e+03 | 3.8634e+03 | 3.2214e+03 | 3.3822e+03 | 3.3787e+03 | 3.4611e+03 | 3.7220e+03 | 3.6973e+03 | 3.6789e+03 | 3.7446e+03 |
| 861 | 3037e+03 | 4.2048e+03 | 4.3122e+03 | 4.3592e+03 | 4.4107e+03 | 3.8702e+03 | 4.0334e+03 | 4.0236e+03 | 4.1241e+03 | 4.2523e+03 | 4.2896e+03 | 4.2976e+03 | 4.3815e+03 |
| 862 | 2260e+03 | 5.1632e+03 | 5.2552e+03 | 5.2695e+03 | 5.3657e+03 | 4.7278e+03 | 4.9137e+03 | 4.8960e+03 | 5.0216e+03 | 5.2136e+03 | 5.2353e+03 | 5.2011e+03 | 5.3271e+03 |
| 863 | 6145e+03 | 6.3829e+03 | 6.5974e+03 | 6.6246e+03 | 6.7604e+03 | 5.8396e+03 | 6.0945e+03 | 6.0464e+03 | 6.2884e+03 | 6.4894e+03 | 6.5713e+03 | 6.4321e+03 | 6.7007e+03 |
| 864 | 1395e+03 | 2.0674e+03 | 2.1423e+03 | 2.1453e+03 | 2.1970e+03 | 1.9499e+03 | 2.0481e+03 | 2.0470e+03 | 2.1031e+03 | 2.1090e+03 | 2.1498e+03 | 2.1881e+03 | 2.2084e+03 |
| 865 | 5929e+03 | 4.5240e+03 | 4.6165e+03 | 4.6592e+03 | 4.7310e+03 | 4.1294e+03 | 4.3082e+03 | 4.2967e+03 | 4.4022e+03 | 4.5848e+03 | 4.6100e+03 | 4.6129e+03 | 4.6903e+03 |
| 866 | 5130e+03 | 4.4185e+03 | 4.5017e+03 | 4.5610e+03 | 4.6028e+03 | 4.0345e+03 | 4.2039e+03 | 4.1953e+03 | 4.2949e+03 | 4.4836e+03 | 4.5121e+03 | 4.4890e+03 | 4.5838e+03 |
| 867 | 1020e+03 | 5.0112e+03 | 5.1478e+03 | 5.1282e+03 | 5.2448e+03 | 4.6596e+03 | 4.8532e+03 | 4.8314e+03 | 4.9652e+03 | 5.0716e+03 | 5.1190e+03 | 5.1079e+03 | 5.2277e+03 |
| 868 | 0985e+03 | 3.9216e+03 | 4.0751e+03 | 4.1158e+03 | 4.2135e+03 | 3.5038e+03 | 3.7353e+03 | 3.7254e+03 | 3.8698e+03 | 4.0236e+03 | 4.0929e+03 | 4.0086e+03 | 4.1791e+03 |
| 869 | | | | | | | | | | | | | |

*Fuente:* Elaborado por el autor

### 3.3.2. Etiquetado de audios de emociones

Una vez obtenidas las características del audio de emociones, se procede a relacionar los datos obtenidos con la emoción asociada, lo cual se le conoce como el proceso de etiquetado de los audios. Para lo cual, se sigue una lógica binaria que se muestra en la Tabla 7 es decir que para cada vector de características extraídas se asocia con un vector binario según corresponda. Este proceso se realiza para todos los audios disponibles en la base de datos.

**Tabla 7**

*Lógica para detección de emociones*

| Ira | Disgusto | Miedo | Felicidad | Neutral | Tristeza |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 |

*Fuente:* Elaborado por el autor

## 3.4. Preparación de datos de entrenamiento

Una vez realizado el procesamiento de audios de emociones, así como, la identificación y extracción de las características de las señales, se procede a crear tres matrices. Para lo cual, se divide la matriz resultante total, del proceso de identificación y extracción de características, en tres grupos, según los porcentajes, 80%, 10% y 10%, que corresponden a los datos de entrenamiento, validación y testeo, respectivamente.

### 3.4.1. Método de normalización

El proceso de normalizar hace que el conjunto de datos se ajuste a valores de escala diferentes a los originales, en el data set que se obtiene se realiza el proceso de normalización, este proceso se basa en establecer un rango determinado al conjunto de datos para que posterior a esto ingrese a la red neuronal y se entrenen de una forma que la red entienda los datos que están ingresando, en el conjunto de datos de entrada se tiene valores máximos y mínimos, a

estos datos se le aplica una normalización establecida en un rango [0,1], los datos negativos se tomaran como los mínimos y los valores más altos como los máximos, gracias a la ecuación (5) se puede normalizar los datos en el rango deseado.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5}$$

En donde:

$X'$es el valor normalizado de cada dato que está en el data set.

$X$es el valor del dato que deseamos normalizar

$X_{min}$ es el valor del dato mínimo que se encuentra en el data set

$X_{max}$ es el valor del dato máximo que se encuentra en el data set

Para el data set que se preparó se establece la normalización en los datos que lo constituyen, por consecuente podemos ver estos datos en la Figura 43.

**Figura 43.**

*Base de datos normalizada con rango [0,1]*

carNorma

868x255 double

| | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 856 | .2143 | 0.2117 | 0.2118 | 0.2149 | 0.2132 | 0.2169 | 0.2180 | 0.2168 | 0.2171 | 0.2114 | 0.2126 | 0.2176 | 0.2154 |
| 857 | .2262 | 0.2241 | 0.2277 | 0.2239 | 0.2264 | 0.2272 | 0.2307 | 0.2316 | 0.2328 | 0.2251 | 0.2288 | 0.2338 | 0.2321 |
| 858 | .2221 | 0.2240 | 0.2254 | 0.2197 | 0.2229 | 0.2295 | 0.2319 | 0.2311 | 0.2319 | 0.2270 | 0.2288 | 0.2318 | 0.2304 |
| 859 | .1875 | 0.1845 | 0.1859 | 0.1861 | 0.1872 | 0.1896 | 0.1931 | 0.1942 | 0.1936 | 0.1862 | 0.1893 | 0.1969 | 0.1931 |
| 860 | .1820 | 0.1870 | 0.1841 | 0.1840 | 0.1847 | 0.1767 | 0.1804 | 0.1816 | 0.1802 | 0.1861 | 0.1839 | 0.1881 | 0.1843 |
| 861 | .2228 | 0.2219 | 0.2229 | 0.2229 | 0.2217 | 0.2275 | 0.2299 | 0.2305 | 0.2295 | 0.2232 | 0.2252 | 0.2319 | 0.2282 |
| 862 | .2870 | 0.2895 | 0.2883 | 0.2851 | 0.2863 | 0.2947 | 0.2967 | 0.2967 | 0.2961 | 0.2905 | 0.2912 | 0.2959 | 0.2933 |
| 863 | .3837 | 0.3755 | 0.3814 | 0.3776 | 0.3806 | 0.3818 | 0.3864 | 0.3840 | 0.3901 | 0.3798 | 0.3845 | 0.3831 | 0.3880 |
| 864 | .0720 | 0.0712 | 0.0725 | 0.0716 | 0.0721 | 0.0771 | 0.0791 | 0.0806 | 0.0794 | 0.0732 | 0.0758 | 0.0825 | 0.0785 |
| 865 | .2429 | 0.2444 | 0.2440 | 0.2434 | 0.2434 | 0.2478 | 0.2508 | 0.2512 | 0.2501 | 0.2465 | 0.2476 | 0.2542 | 0.2495 |
| 866 | .2374 | 0.2370 | 0.2361 | 0.2366 | 0.2347 | 0.2404 | 0.2428 | 0.2436 | 0.2421 | 0.2394 | 0.2407 | 0.2455 | 0.2421 |
| 867 | .2784 | 0.2788 | 0.2809 | 0.2754 | 0.2781 | 0.2894 | 0.2922 | 0.2918 | 0.2919 | 0.2806 | 0.2831 | 0.2893 | 0.2865 |
| 868 | .2085 | 0.2020 | 0.2065 | 0.2062 | 0.2084 | 0.1988 | 0.2072 | 0.2079 | 0.2106 | 0.2072 | 0.2115 | 0.2114 | 0.2143 |
| 869 | | | | | | | | | | | | | |

*Fuente:* Elaborado por el autor

### 3.5. Entrenamiento de la red neuronal

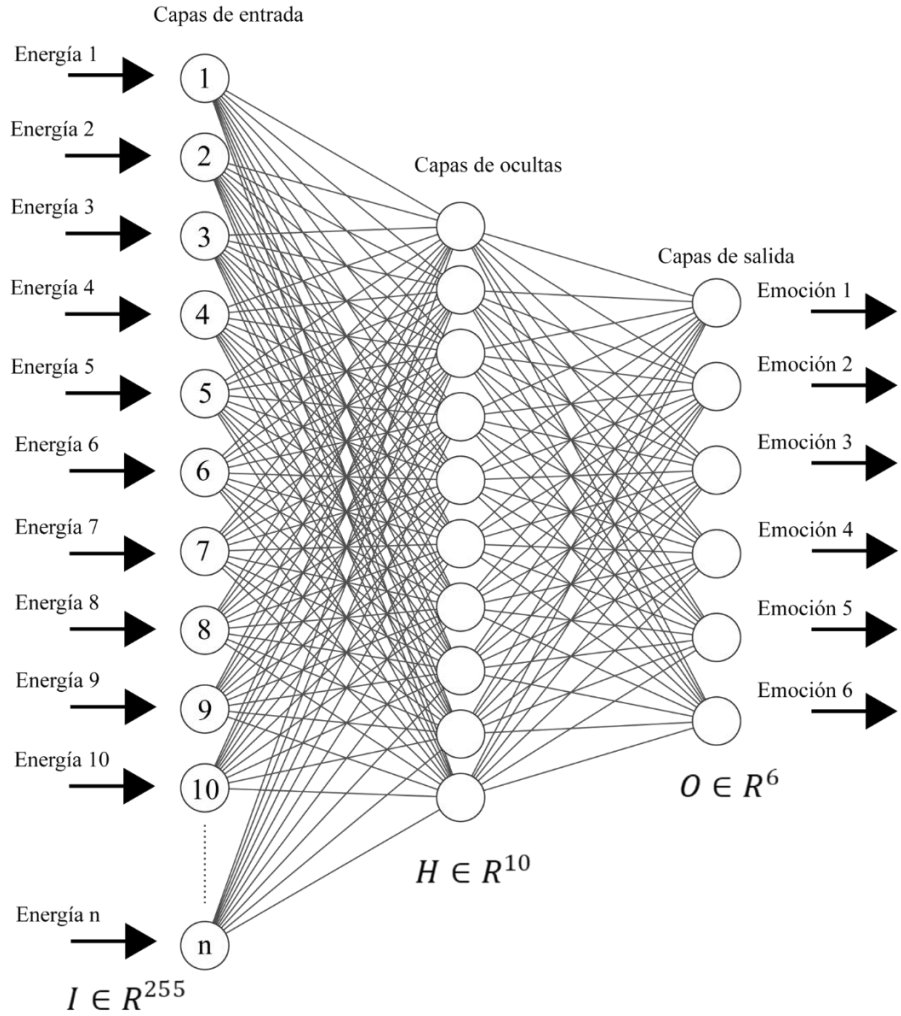
Este es un proceso de tres etapas, que son: creación del modelo LSTM, parametrización y entrenamiento de la red neuronal.

#### 3.5.1. Creación del modelo LSTM

La arquitectura del modelo LSTM se crea mediante la aplicación Deep Network Designer de MATLAB, en esta se establecen dos modelos de los cuales se realizan las respectivas pruebas para determinar el modelo más optimo en la predicción de emociones. En la Figura 44 se observa la arquitectura general que se adopta para la red neuronal se parte con un número de neuronas en la capa de entrada de 255 y 6 neuronas en la capa de salida.

**Figura 44.**

*Arquitectura de red neuronal para clasificación de emociones*




*Nota.* Adaptado de (Calin, 2020)

#### *3.5.1.1. Cálculo de neuronas en las capas ocultas*

El número de neuronas ocultas en cada capa depende del modelo que se vaya a utilizar, existen recomendaciones o fórmulas para realizar el cálculo de este número de neuronas, pero no es un estándar, pues cada problema que se requiera solucionar con redes neuronales tiene su nivel de complejidad y no es una fórmula general para el cálculo de estas, sin embargo, es un gran aporte ya que brinda un número aproximado para empezar a realizar el cálculo de las neuronas.

De esta manera se toma en cuenta la fórmula establecida por el canal (Xpikuos, 2019). En donde en la ecuación 6 se establece la fórmula para la obtención del número de neuronas para una capa oculta, mientras que en la ecuación 7 se determina el número de neuronas para dos capas, con ello se hace el cálculo con los datos de esta investigación que es de 255 neuronas de entrada y 6 de salida.

$$h = (in * out)^{\frac{1}{2}} \tag{6}$$

donde;

$h$ = número de neuronas en la capa oculta

$in$ = número de neuronas en la capa de entrada

$out$ = número de neuronas en la capa de salida

Aplicando los datos que se tiene para la presente investigación el número de capas queda de la siguiente manera:

$$h = (255 * 6)^{\frac{1}{2}}$$

$$\boldsymbol{h = 39.11}$$

El resultado de las neuronas de la primera capa es de aproximadamente 40, en la siguiente sección se realiza pruebas con este número de neuronas para ver la efectividad de dicha formula en este caso de estudio.

Ahora se hace el cálculo para dos capas ocultas en nuestra red neuronal.

$$h1 = out * r^2$$
$$h2 = out * r \quad (7)$$
$$r = \left(\frac{in}{out}\right)^{\frac{1}{3}}$$

Donde;

*h1* = número de neuronas en la capa oculta 1

*h2* = número de neuronas en la capa oculta 2

*in* = número de neuronas en la capa de entrada

*out* = número de neuronas en la capa de salida

El cálculo de neuronas para dos capas quedaría de la siguiente forma:

$$r = \left(\frac{255}{6}\right)^{\frac{1}{3}}$$

$$r = 3.44$$

$$h1 = 6 * 11.9$$

$$\boldsymbol{h1 = 71.4\ neuronas}$$

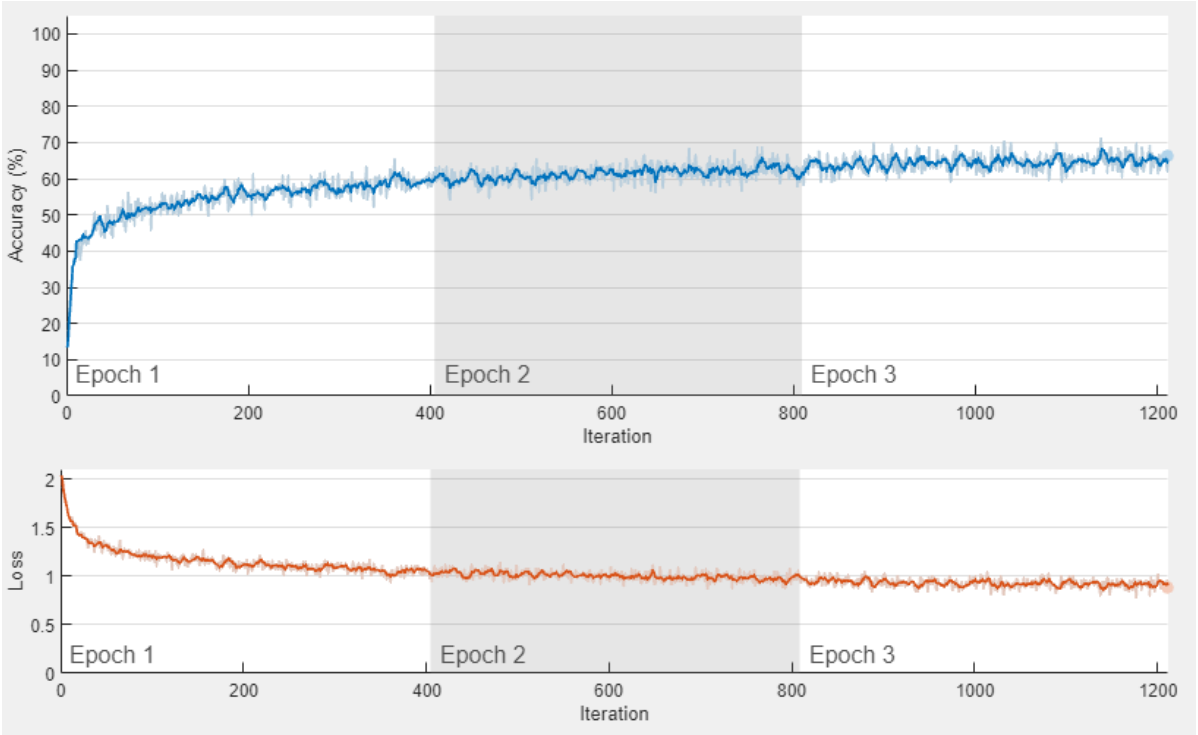
$$h2 = 6 * 3.44$$

$$\boldsymbol{h2 = 20.64\ neuronas}$$

Con los números de neuronas obtenidos se realiza la prueba para ver cómo se efectúa el entrenamiento. En la Figura 45 se observa que se tiene un nivel de accuracy muy bajo con el número de 40 neuronas, por lo que no es óptimo usar esta técnica en esta investigación, también la pérdida es muy elevada por lo que se estaría causando Underfitting, el cual es un aprendizaje ineficiente debido al bajo número de neuronas.

**Figura 45.**

*Entrenamiento realizado para obtener la precisión y el error con 40 neuronas en la primera capa oculta*



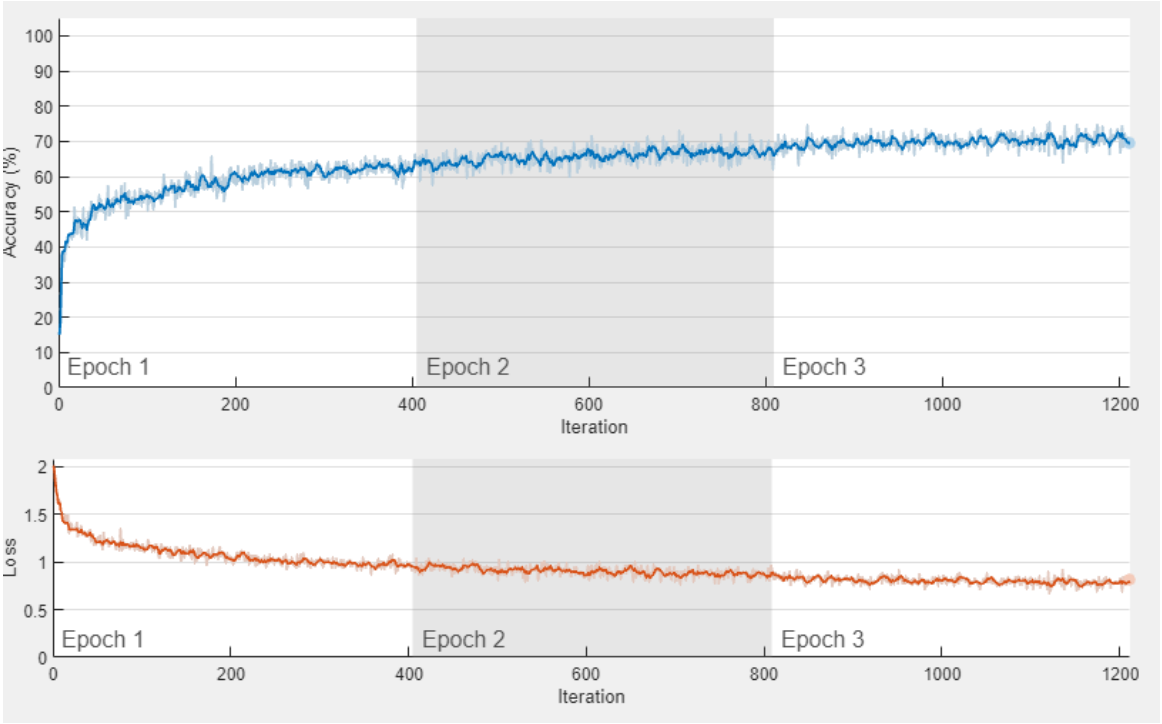

*Fuente:* Elaborado por el autor

Para las capas conformadas por 72 y 21 neuronas se aprecia que no existe un incremento en el accuracy del entrenamiento, de igual forma que la anterior iteración se tiene un número muy alto para el error como se observa en la Figura 46.

Es recomendable aumentar capas para disminuir el Underfitting, tomando en cuenta que no se debe incrementar muchas capas pues ocasionaría Overfitting lo que es el sobre entrenamiento de la red y por ende la red neuronal no sería capaz de realizar predicciones correctas.

**Figura 46.**

*Entrenamiento realizado para obtener la precisión y el error con 72 neuronas en la primera capa oculta y 21 en la segunda capa oculta*




*Fuente:* Elaborado por el autor

### *3.5.1.2. Modelos de redes neuronales LSTM*

Se presentan dos modelos de redes neuronales los cuales se van a estudiar y analizar con el fin de seleccionar el óptimo para la presente investigación, en la Tabla 8 se presenta las neuronas de entrada y salida, el número de capas, y el tipo de modelo que se usa para realizar el entrenamiento.

**Tabla 8**

*Modelos propuestos para el entrenamiento*

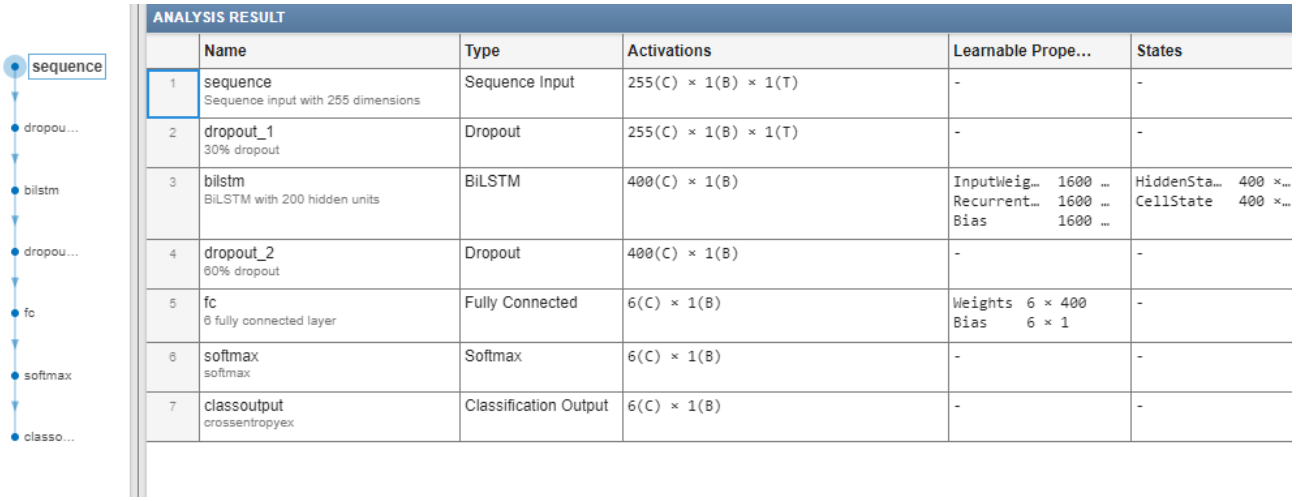
| | **Modelo 1** | **Modelo 2** |
|---|---|---|
| Input | 255 | 255 |
| Outputs | 6 | 6 |
| Capas | 7 | 8 |
| Tipo | BiLSTM | LSTM |

*Fuente:* Elaborado por el autor

En la Figura 47 se muestra el modelo 1 que se utilizó en el entrenamiento de la red neuronal.

**Figura 47.**

*Arquitectura de red neuronal BiLSTM con 7 capas para la clasificación de emociones*




ANALYSIS RESULT

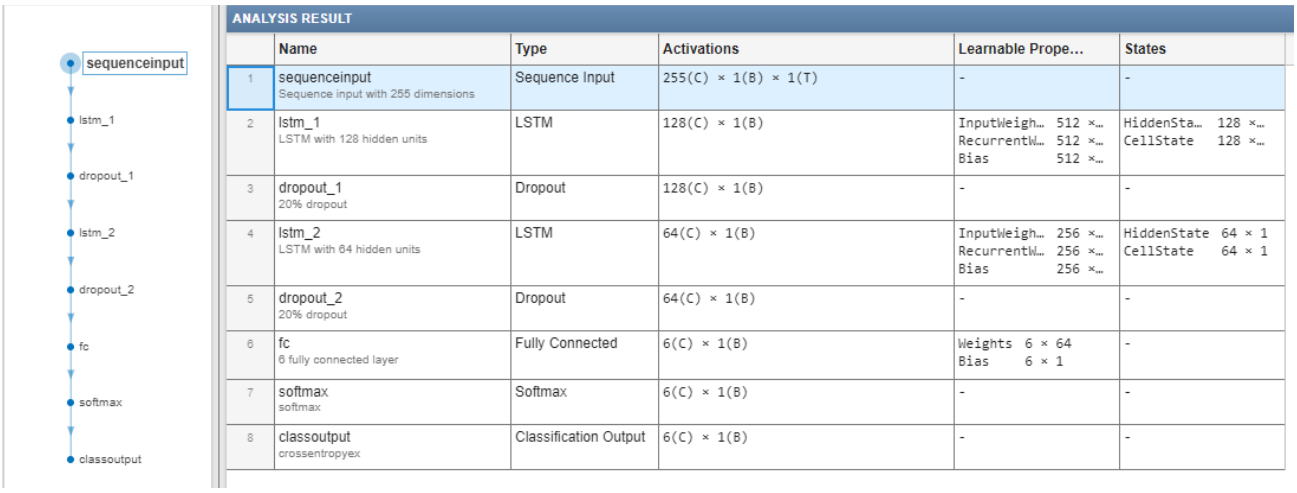
| | Name | Type | Activations | Learnable Prope... | States |
|---|---|---|---|---|---|
| 1 | sequence<br>Sequence input with 255 dimensions | Sequence Input | 255(C) × 1(B) × 1(T) | - | - |
| 2 | dropout_1<br>30% dropout | Dropout | 255(C) × 1(B) × 1(T) | - | - |
| 3 | bilstm<br>BiLSTM with 200 hidden units | BiLSTM | 400(C) × 1(B) | InputWeig… 1600 …<br>Recurrent… 1600 …<br>Bias 1600 … | HiddenSta… 400 ×…<br>CellState 400 ×… |
| 4 | dropout_2<br>60% dropout | Dropout | 400(C) × 1(B) | - | - |
| 5 | fc<br>6 fully connected layer | Fully Connected | 6(C) × 1(B) | Weights 6 × 400<br>Bias 6 × 1 | - |
| 6 | softmax<br>softmax | Softmax | 6(C) × 1(B) | - | - |
| 7 | classoutput<br>crossentropyex | Classification Output | 6(C) × 1(B) | - | - |

*Fuente:* Elaborado por el autor

En la Figura 48 se indica la arquitectura del modelo 2, el cual consta de 8 capas.

**Figura 48.**

*Arquitectura de red neuronal LSTM con 8 capas para la clasificación de emociones*




ANALYSIS RESULT

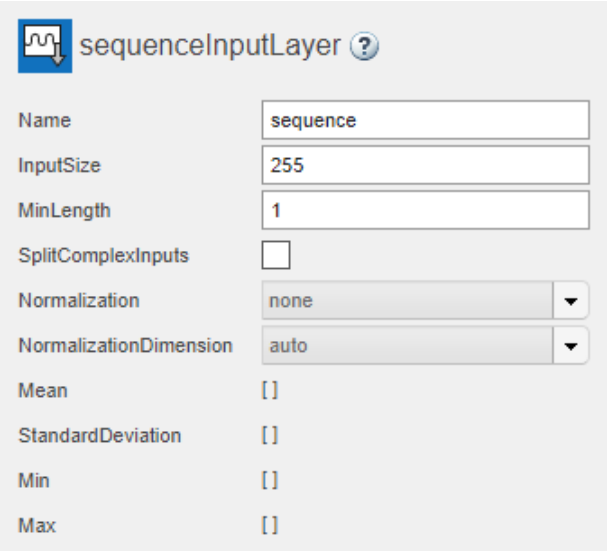
| | Name | Type | Activations | Learnable Prope... | States |
|---|---|---|---|---|---|
| 1 | sequenceinput<br>Sequence input with 255 dimensions | Sequence Input | 255(C) × 1(B) × 1(T) | - | - |
| 2 | lstm_1<br>LSTM with 128 hidden units | LSTM | 128(C) × 1(B) | InputWeigh… 512 ×…<br>RecurrentW… 512 ×…<br>Bias 512 ×… | HiddenSta… 128 ×…<br>CellState 128 ×… |
| 3 | dropout_1<br>20% dropout | Dropout | 128(C) × 1(B) | - | - |
| 4 | lstm_2<br>LSTM with 64 hidden units | LSTM | 64(C) × 1(B) | InputWeigh… 256 ×…<br>RecurrentW… 256 ×…<br>Bias 256 ×… | HiddenState 64 × 1<br>CellState 64 × 1 |
| 5 | dropout_2<br>20% dropout | Dropout | 64(C) × 1(B) | - | - |
| 6 | fc<br>6 fully connected layer | Fully Connected | 6(C) × 1(B) | Weights 6 × 64<br>Bias 6 × 1 | - |
| 7 | softmax<br>softmax | Softmax | 6(C) × 1(B) | - | - |
| 8 | classoutput<br>crossentropyex | Classification Output | 6(C) × 1(B) | - | - |

*Fuente:* Elaborado por el autor

Los modelos de redes neuronales presentados constan de varias capas, estas son: capa de entrada, capas intermedias y capas de salida. Un aspecto importante, que debe considerarse es que la capa de entrada se activa con un conjunto de datos de dimensiones 255 (C) x 1 (B) x 1 (T), como se muestra en la Figura 49, estas dimensiones corresponden a las características obtenidas mediante la transformada de Wavelet, esto implica que un audio con una transformada multinivel de Wavelet diferente, hará que los resultados en la clasificación sean erróneos.

**Figura 49.**

*Capa de secuencia de entrada*




*Fuente:* Elaborado por el autor

En la Figura 50 se observa la matriz que ingresa a la red neuronal para ser entrenada, el número de filas es de 868, estas corresponden a cada emoción que se tiene en la base de datos y el número de columnas es de 255 que pertenece a la energía de cada señal extraída de la descomposición multinivel de Wavelet. La red neuronal toma de entrada cada vector de 255 elementos y lo asimila como una emoción para dicho entrenamiento.

**Figura 50.**

*Conjunto de datos de entrenamiento*

car

868x255 double

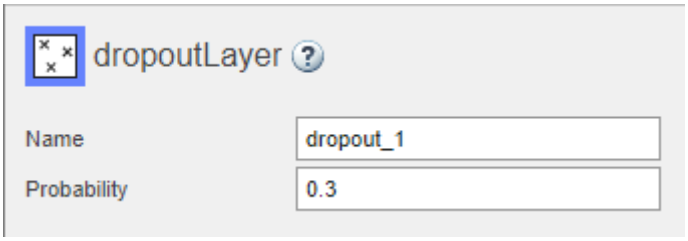
| | 243 | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 857 | 3529e+03 | 4.2353e+03 | 4.3810e+03 | 4.3746e+03 | 4.4803e+03 | 3.8652e+03 | 4.0442e+03 | 4.0373e+03 | 4.1688e+03 | 4.2788e+03 | 4.3410e+03 | 4.3242e+03 | 4.4378e+03 |
| 858 | 2943e+03 | 4.2348e+03 | 4.3476e+03 | 4.3128e+03 | 4.4273e+03 | 3.8950e+03 | 4.0594e+03 | 4.0313e+03 | 4.1566e+03 | 4.3065e+03 | 4.3407e+03 | 4.2967e+03 | 4.4125e+03 |
| 859 | 7978e+03 | 3.6746e+03 | 3.7781e+03 | 3.8207e+03 | 3.9006e+03 | 3.3853e+03 | 3.5496e+03 | 3.5450e+03 | 3.6413e+03 | 3.7235e+03 | 3.7750e+03 | 3.8029e+03 | 3.8714e+03 |
| 860 | 7189e+03 | 3.7097e+03 | 3.7518e+03 | 3.7907e+03 | 3.8634e+03 | 3.2214e+03 | 3.3822e+03 | 3.3787e+03 | 3.4611e+03 | 3.7220e+03 | 3.6973e+03 | 3.6789e+03 | 3.7446e+03 |
| 861 | 3037e+03 | 4.2048e+03 | 4.3122e+03 | 4.3592e+03 | 4.4107e+03 | 3.8702e+03 | 4.0334e+03 | 4.0236e+03 | 4.1241e+03 | 4.2523e+03 | 4.2896e+03 | 4.2976e+03 | 4.3815e+03 |
| 862 | 2260e+03 | 5.1632e+03 | 5.2552e+03 | 5.2695e+03 | 5.3657e+03 | 4.7278e+03 | 4.9137e+03 | 4.8960e+03 | 5.0216e+03 | 5.2136e+03 | 5.2353e+03 | 5.2011e+03 | 5.3271e+03 |
| 863 | 6145e+03 | 6.3829e+03 | 6.5974e+03 | 6.6246e+03 | 6.7604e+03 | 5.8396e+03 | 6.0945e+03 | 6.0464e+03 | 6.2884e+03 | 6.4894e+03 | 6.5713e+03 | 6.4321e+03 | 6.7007e+03 |
| 864 | 1395e+03 | 2.0674e+03 | 2.1423e+03 | 2.1453e+03 | 2.1970e+03 | 1.9499e+03 | 2.0481e+03 | 2.0470e+03 | 2.1031e+03 | 2.1090e+03 | 2.1498e+03 | 2.1881e+03 | 2.2084e+03 |
| 865 | 5929e+03 | 4.5240e+03 | 4.6165e+03 | 4.6592e+03 | 4.7310e+03 | 4.1294e+03 | 4.3082e+03 | 4.2967e+03 | 4.4022e+03 | 4.5848e+03 | 4.6100e+03 | 4.6129e+03 | 4.6903e+03 |
| 866 | 5130e+03 | 4.4185e+03 | 4.5017e+03 | 4.5610e+03 | 4.6028e+03 | 4.0345e+03 | 4.2039e+03 | 4.1953e+03 | 4.2949e+03 | 4.4836e+03 | 4.5121e+03 | 4.4890e+03 | 4.5838e+03 |
| 867 | 1020e+03 | 5.0112e+03 | 5.1478e+03 | 5.1282e+03 | 5.2448e+03 | 4.6596e+03 | 4.8532e+03 | 4.8314e+03 | 4.9652e+03 | 5.0716e+03 | 5.1190e+03 | 5.1079e+03 | 5.2277e+03 |
| 868 | 0985e+03 | 3.9216e+03 | 4.0751e+03 | 4.1158e+03 | 4.2135e+03 | 3.5038e+03 | 3.7353e+03 | 3.7254e+03 | 3.8698e+03 | 4.0236e+03 | 4.0929e+03 | 4.0086e+03 | 4.1791e+03 |
| 869 | | | | | | | | | | | | | |

*Fuente:* Elaborado por el autor

Asimismo, se dispone de las capas intermedias (dropout) que evitan el sobreajuste y permite combinar de forma exponencial muchas arquitecturas de red neuronal diferente, observar Figuras 51 y 52. El término "dropout" se refiere a la eliminación de unidades (ocultas y visibles) en una red neuronal. Por "abandonar" una unidad se entiende como eliminarla temporalmente de la red, junto con todas sus conexiones entrantes y salientes. En este sentido, la elección de las unidades que se eliminan es aleatoria. En el caso más sencillo, cada unidad se retiene con una probabilidad fija p independiente de otras unidades, donde p se fija en 0,3 y 0,6, lo que parece estar cerca de ser óptimo para una amplia gama de redes y tareas.

**Figura 51.**

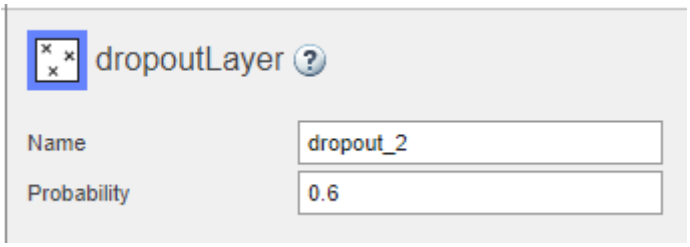
*Capa Dropout con probabilidad de 0.3*




*Fuente:* Elaborado por el autor

**Figura 52.**

*Capa Dropout con probabilidad de 0.6*



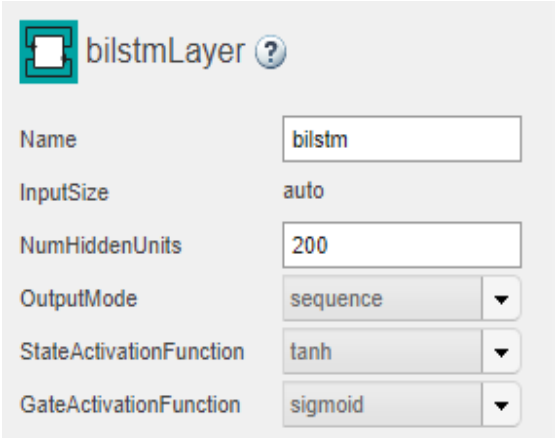

*Fuente:* Elaborado por el autor

En la siguiente capa el número de neuronas establecidas es de 200, pero al ser una arquitectura bidireccional LSTM este número se duplica y se establecen en 400 neuronas, como se muestra en la Figura 53, las conexiones bidireccionales a largo plazo entre unidades temporales de series temporales y datos secuenciales se aprenden mediante una capa LSTM bidireccional (BiLSTM).

Cuando se desea que la red aprenda series temporales completas en cada unidad temporal, estas dependencias pueden resultar útiles. La red BiLSTM decide si aprende o descarta información relevante abriendo y cerrando compuertas, esto con las funciones de activación sigmoidal y tangente hiperbólica, estas compuertas se denominan input gate, forget gate y output gate.

**Figura 53.**

*Capa BiLSTM*

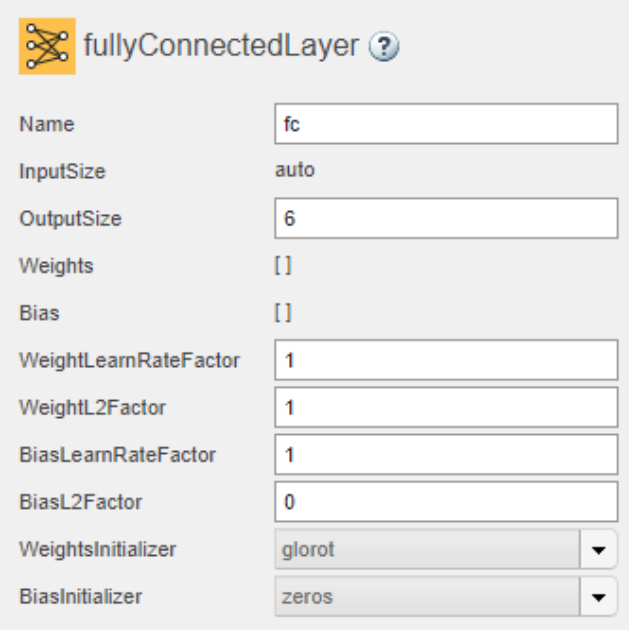



*Fuente:* Elaborado por el autor

Por otro lado, una capa totalmente conectada (fully connected) se encarga de aplicar una transformación lineal al vector de entrada a través de una matriz de pesos, ver Figura 54. Como resultado, están presentes todas las conexiones posibles entre capas, lo que significa que cada entrada del vector de entrada influye en cada salida del vector de salida. Lo que significa que las 400 neuronas de salida de la capa anterior se conectan con las 6 neuronas de la capa fully Connected, el numero pertenece a la clasificación de las emociones que deseamos como resultado.

**Figura 54.**

*Capa Fully Connected*

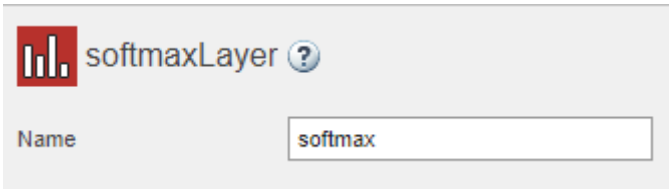



*Fuente:* Elaborado por el autor

La capa softmax es una función que convierte el vector de K valores reales en un vector de K valores reales que suman 1. Los valores de entrada pueden ser positivos, negativos, cero o mayores que uno, no obstante, la capa softmax los transforma en valores entre 0 y 1, de modo que pueden interpretarse como probabilidades. Si una de las entradas es pequeña o negativa, la capa softmax la convierte en una pequeña probabilidad, y si una entrada es grande, entonces la convierte en una probabilidad grande, pero siempre permanecerá entre 0 y 1, ver Figura 55.

**Figura 55.**

*Capa Softmax*




*Fuente:* Elaborado por el autor

Por último, la capa classoutput que se observa en la Figura 56 se encarga de interpretar los arreglos provenientes de la capa fully connected. De esta forma, se obtiene la emoción a la que corresponde según la onda de ingreso en la capa de entrada y al etiquetado que se le estableció en la sección 3.3.2.

**Figura 56.**

*Capa de clasificación*


*Fuente:* Elaborado por el autor

### 3.5.2. Parametrización y entrenamiento

Una vez definidas los modelos de la red neuronal se configura los parámetros de entrenamiento. Para ello se realizan ensayos, estos varían en el número de épocas y en el tamaño mínimo de lote, pero se mantienen constante en el algoritmo y la tasa de aprendizaje.

#### *3.5.2.1. Modelo 1*

Se realiza el entrenamiento con diversos parámetros para ver el mejor funcionamiento de la red, se asignan tamaños mínimos de lotes diferentes y el número de épocas, este proceso

se lo realiza para los dos modelos explicados anteriormente, los parámetros que se emplean para el modelo 1 se detallan en la Tabla 9.

El algoritmo para usar es el descenso del gradiente ya que este ayuda a mejorar el entrenamiento de forma que avanzan las épocas, para disminuir la función de pérdida o de error.

Los números de épocas que constan en la Tabla 9 se consideran a partir de la tercera época en donde tras realizar pruebas se constata que se obtiene un entrenamiento favorable, por ende, se usa 3, 4 y 5 épocas para los ensayos que se realizará.

Para el número mínimo de lote se toma en cuenta el valor total de los datos, en este caso es de 868 que corresponde al número máximo de audios de emociones, de esta manera tomamos valores que dividan este número en pequeños lotes, por lo que se inicia con un lote de 62 que se repetirá 14 veces en las épocas establecidas, consecuentemente se toma el valor de lote de 217 que se repetirá 4 veces en las épocas y finalmente se toma el valor total para el lote de 868 que se repetirá una vez en las épocas.

La taza de aprendizaje se la estandarizó en 0.005 debido a que es un valor intermedio y recomendable para trabajar con el algoritmo del descenso del gradiente, ya que si situamos un valor más alto impediría que el algoritmo converja y con un valor muy bajo el algoritmo tardaría mucho en establecer ponderaciones óptimas.

**Tabla 9**

*Ensayos para el primer modelo.*

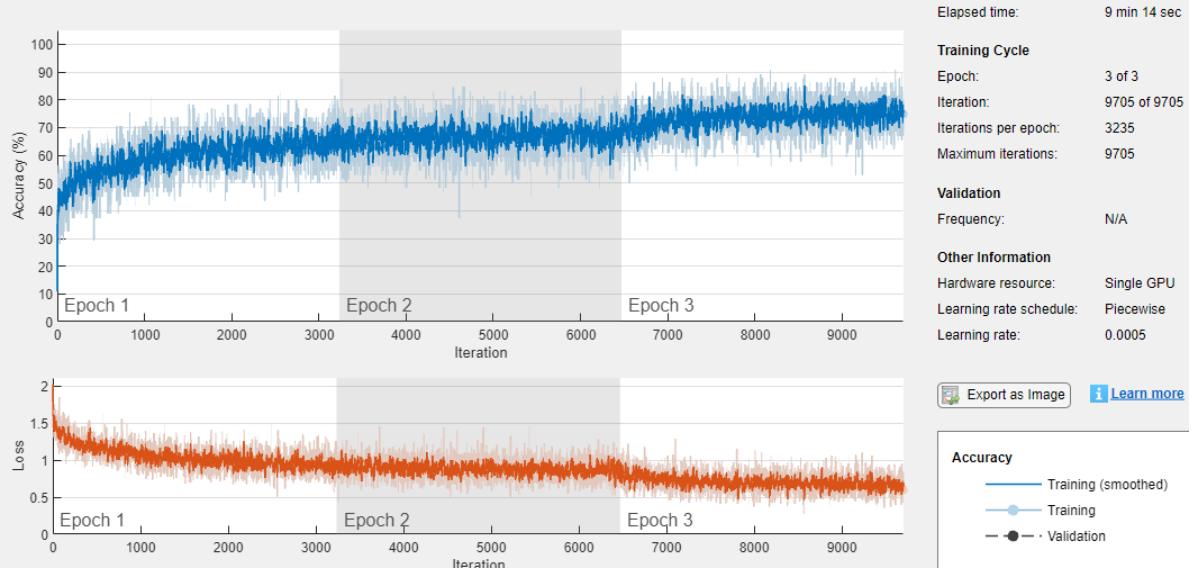
| Parámetro | Ensayo 1 | Ensayo 2 | Ensayo 3 |
|---|---|---|---|
| Algoritmo | Descenso del gradiente | Descenso del gradiente | Descenso del gradiente |
| Número de épocas | 3 | 4 | 5 |
| Tamaño mínimo de lote | 62 | 217 | 868 |
| Tasa de aprendizaje | 0.005 | 0.005 | 0.005 |

*Fuente:* Elaborado por el autor

Se observa en la Figura 57 el resultado del primer ensayo con los datos de parametrización establecidos en la Tabla 9, el sobre entrenamiento de la red es notorio en la gráfica, ya que se tiene tamaños muy pequeños en el lote y por ende no es adecuado realizar el sistema con este tipo de parámetros. Asimismo, el entrenamiento dura un tiempo considerablemente alto y no se obtienen resultados deseados ya que las pérdidas son muy altas y el accuracy muy bajo.

**Figura 57.**

*Resultado de la precisión y error del primer ensayo aplicando el modelo 1*

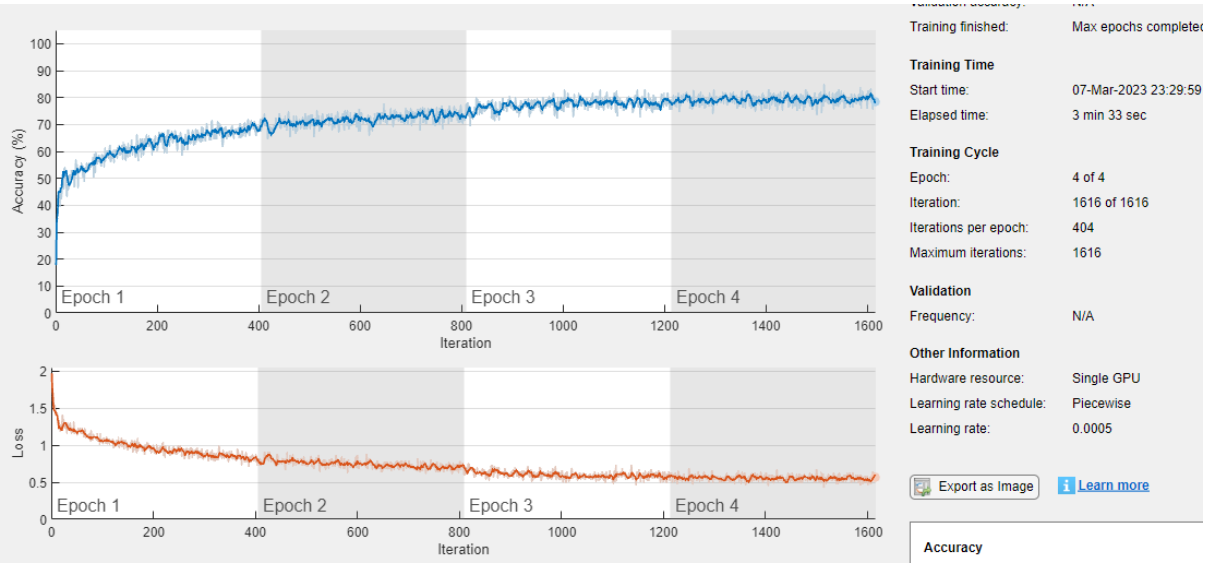



*Fuente:* Elaborado por el autor

A continuación, en la Figura 58 se realiza el entrenamiento de la red con los parámetros establecidos en el ensayo 2 de la tabla 9, el cual se obtiene que el entrenamiento es estable, sin llegar a un sobreajuste en la red, la cuarta época es un innecesaria ya que se observa que desde la época 2 se mantiene estable en accuracy y en los. También el tiempo de ejecución del entrenamiento es más corto, esto depende de las características del CPU ya que influye bastante al momento de realizar el entrenamiento.

**Figura 58.**

*Resultado de la precisión y error del segundo ensayo aplicando el modelo 1*

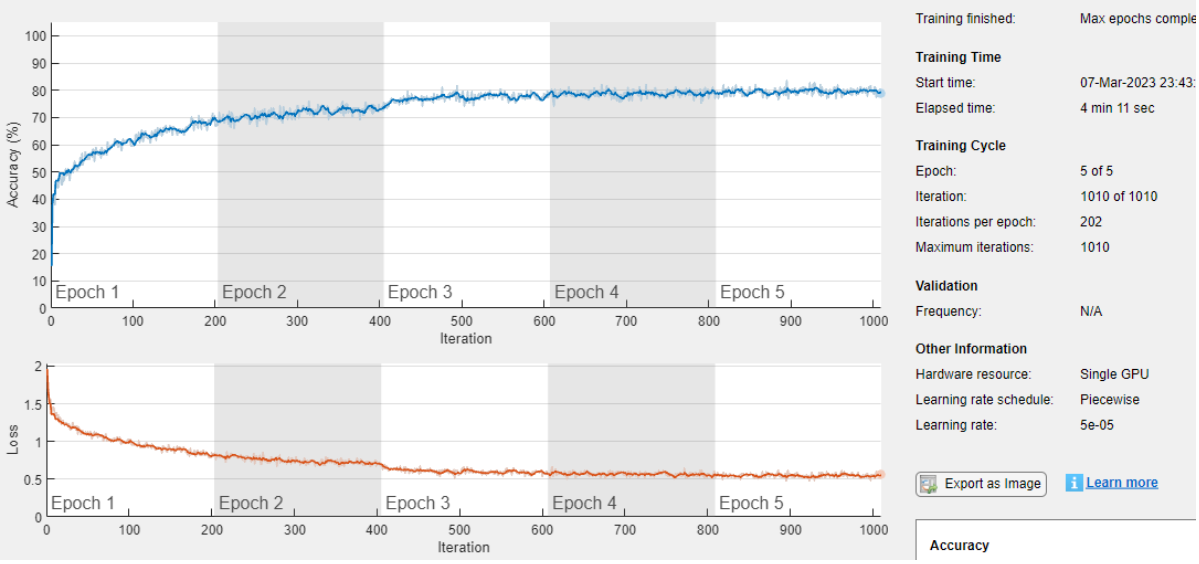



*Fuente:* Elaborado por el autor

Para el último ensayo que se tiene en la Figura 59 se obtienen resultados similares al ensayo 2 con mayor número de épocas y mayor numero en el tamaño de lote, por lo que el tiempo de entrenamiento es más alto que el ensayo 2, asimismo, a partir de la época 3 se estabiliza el entrenamiento y no es necesario más épocas.

**Figura 59.**

*Resultado de la precisión y error del tercer ensayo aplicando el modelo 1*




*Fuente:* Elaborado por el autor

Durante el entrenamiento de la red neuronal se verifica que la tasa de aprendizaje fijada en 0.005 es más estable que, la tasa fijada en 0.01. En consecuencia, la perdida de mínima de lotes se mantuvo estable en los ensayos 2 y 3, también se puede ver que el ensayo 2 es el óptimo para nuestro entrenamiento ya que el tamaño de lote es de 217 y se divide el conjunto de datos en múltiplos del total, ya que no es conveniente tomar todos los datos e ingresarlos a la red para el entrenamiento.

En contraste, la pérdida de lotes en la primera iteración es alta, mientras que, a medida que las iteraciones avanzan, las pérdidas disminuyen de forma considerable. Este resultado es favorable, debido a que, se espera que al término de las iteraciones las pérdidas de paquetes sean mínimas.

#### *3.5.2.2. Modelo 2*

Para el segundo modelo realizamos el mismo número de ensayos que se realizó con el modelo 1, para ello establecemos los valores de parametrización en la Tabla 10, en donde se hace el cambio de los parámetros del tamaño mínimo de lote y el número de épocas para cada ensayo, para el número mínimo de lote se toma en cuenta el valor total de los datos, en este caso es de 868 que corresponde al número máximo de audios de emociones, se obtiene múltiplos de este valor y se toma un valor bajo medio y el más alto.

**Tabla 10**

Ensayos para el segundo modelo.

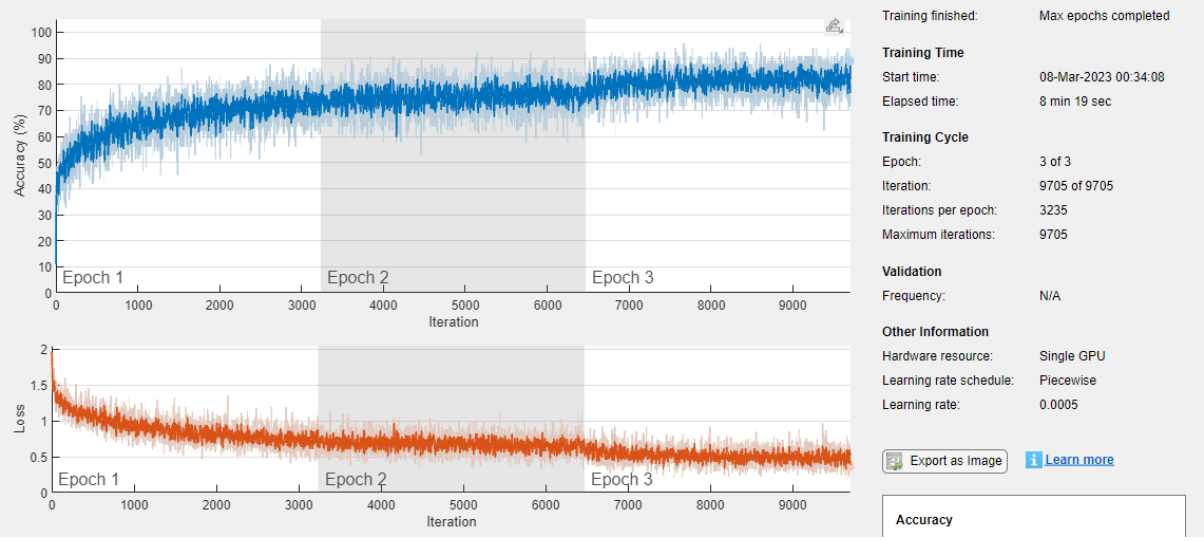
| Parámetro | Ensayo 1 | Ensayo 2 | Ensayo 3 |
|---|---|---|---|
| Algoritmo | Descenso del gradiente | Descenso del gradiente | Descenso del gradiente |
| Numero de épocas | 3 | 4 | 5 |
| Tamaño mínimo de lote | 62 | 217 | 868 |
| Tasa de aprendizaje | 0.005 | 0.005 | 0.005 |

*Fuente:* Elaborado por el autor

Para el primer ensayo con los datos de parametrización que están establecidos en la tabla 9 se observa en la Figura 60 que el entrenamiento tiene un sobre ajuste por lo cual no es óptimo, esto es visible en la exactitud del entrenamiento y en la perdida, ya que tienen muchas fluctuaciones y no se estabiliza de forma constante la función de error. Asimismo, al tener un número muy pequeño en el tamaño de lote el entrenamiento tarda bastante tiempo.

**Figura 60.**

*Resultado de la precisión y error del primer ensayo aplicando el modelo 2*

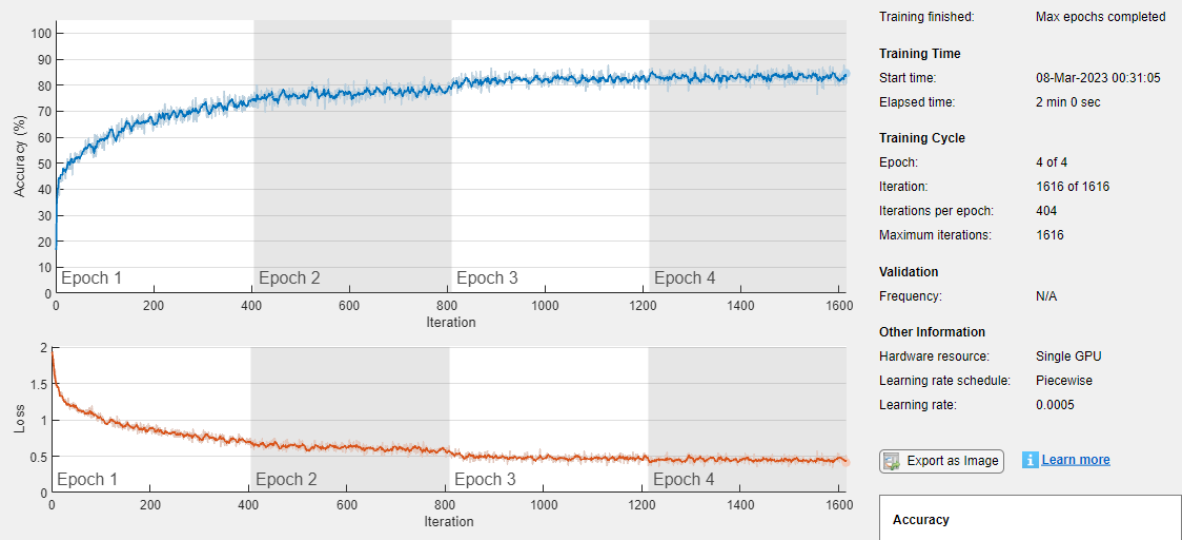



*Fuente:* Elaborado por el autor

En el segundo ensayo que se tiene en la Figura 61, se aprecia la estabilidad en el entrenamiento, ya que se tiene un numero de lote intermedio, la exactitud del modelo y las pérdidas son estables.

**Figura 61.**

*Resultado de la precisión y error del segundo ensayo aplicando el modelo 2*

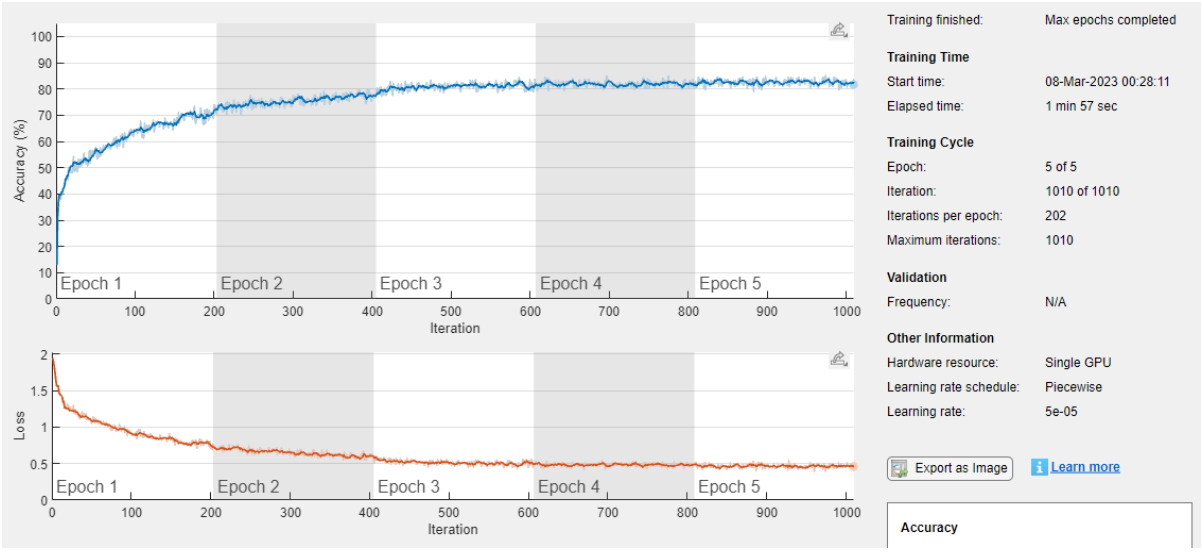



*Fuente:* Elaborado por el autor

Al final se tiene el entrenamiento para el ensayo 3 que se aprecia en la Figura 62, en el cual se obtiene resultados óptimos a partir de la época 3 en donde los valores de exactitud y perdida se estabilizan, pero no es recomendable usar el tamaño total de datos en el tamaño de lote.

**Figura 62.**

*Resultado de la precisión y error del tercer ensayo aplicando el modelo 2*




*Fuente:* Elaborado por el autor

En contraste, el modelo 1 y el modelo 2 que se establecieron tomaron resultados parecidos y por ende los datos de parametrización se establece en la Tabla 11, en donde son los óptimos para el entrenamiento, el modelo que se aplicara en el sistema se lo escoge con los resultados que arroje la matriz de confusión, ya que en esta matriz se detallan los aciertos y errores de la red entrenada.

**Tabla 11**

*Parámetros de entrenamiento seleccionados para la red neuronal*

| **Parámetro** | **Valor** |
|---|---|
| Algoritmo | Descenso del gradiente |
| Numero de épocas | 3 |
| Tamaño mínimo de lote | 217 |
| Tasa de aprendizaje | 0.005 |

*Fuente:* Elaborado por el autor

# CAPITULO 4

Las pruebas de rendimiento del sistema de reconocimiento de emociones por voz se resumen en dos fases en esta sección. La primera es para examinar la proporción de eficacia y el porcentaje de inexactitud para cada emoción considerada, se proporciona en primer lugar la matriz de confusión de cada modelo propuesto en el capítulo 3. Por otra parte, se realizan pruebas aleatorias a seis individuos por separado para determinar el diagnóstico de depresión (leve y grave) de cada paciente.

El aporte brindado por esta investigación se basa en varios factores, el primero es el aporte teórico que fundamenta la investigación, en donde se establecen temas relacionados con la representación del sonido en el habla, la caracterización de estos sonidos, las redes neuronales que existen dentro del aprendizaje profundo la relación de estas con el reconocimiento de emociones y la forma tradicional de detectar la depresión.

Otro de los aportes brindados es la forma de caracterización de las señales, en este sentido, se hace uso de la descomposición multinivel de Wavelet que lo que hace asignar un valor numérico a una señal, este valor depende del nivel de descomposición, en este caso es nivel 7, el cual arroja 255 características en la señal de cada emoción propuesta en la base de datos.

Finalmente se propone una relación entre los resultados obtenidos por el sistema de reconocimiento de emociones y el test que se aplicó a los pacientes para determinar en qué grado de depresión se encuentran, esto sirve de ayuda en al área de psicología para tratar esta patología en dichos pacientes.

## 4.1. Fase de resultados del entrenamiento de la red neuronal

En esta fase veremos la matriz de confusión para los dos modelos planteados, el modelo que tenga un alto nivel de precisión será el que se use para realizar las pruebas de funcionamiento del sistema en pacientes.

### 4.1.1. Modelo 1

Para la evaluación de la técnica propuesta en esta investigación para el reconocimiento de emociones por voz, se muestra la matriz de confusión del primer modelo presentado, mirar Figura 63. De esta forma, se analiza la proporción de eficacia y el porcentaje de inexactitud para cada emoción considerada en los resultados experimentales.

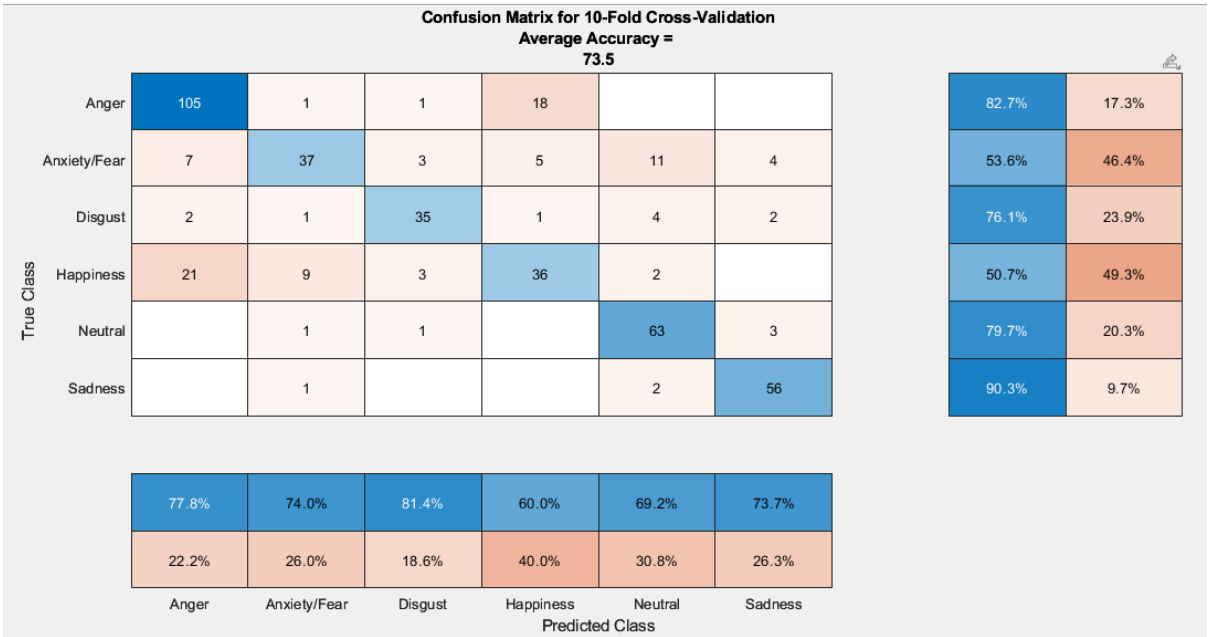
Para el primer caso se tiene un accuracy del 73.5% lo que es relativamente bajo para el desarrollo de sistemas con redes neuronales, sin embargo, esto se debe a la falta de audios en la base de datos, ya que no se cuenta con muchas muestras para que la red se entrene de mejor manera, también influye el modelo escogido y la forma que se extrajo las características de la señal.

El área de interés en este campo es relativamente nuevo pues sus primeras investigaciones surgen a partir del año 2016, en donde (Ganapathy, 2016), muestra en su investigación la recopilación de varios estudios relacionados al reconocimiento de emociones mediante aprendizaje profundo. No obstante, la mayor parte de las investigaciones se desarrollan a partir del año 2018 hasta la actualidad.

En estudios similares se menciona que trabajar con el reconocimiento de emociones es complejo ya que no todas las personas expresan los sentimientos de la misma manera, en este sentido (Aggarwal et al., 2022) en su investigación tiene como resultado un 56.71% de efectividad de la red neuronal que usaron, asimismo, la autora (Sandoval, 2019) en su investigación obtiene resultados entre el 70% y 75% , cabe mencionar que los algoritmos usados y las bases de datos son diferentes a las de esta investigación, pero se tiene la similitud en el porcentaje de precisión.

**Figura 63.**

*Matriz de confusión del Modelo 1*




*Fuente:* Elaborado por el autor

En contraste, la efectividad para la emoción de ira es de 82.7%.30 % con una tasa de error de 17.3%. Asimismo, el disgusto presenta un 76.1 % de eficacia, junto con un 23.9% de clasificaciones erróneas. Del mismo modo, el miedo representa un 53.6 % de detecciones correctas, mientras que, el 46.4 % se asocia con una categoría errónea. De manera similar, la felicidad posee una eficiencia de 50.7 %, con un 49.3 % de errores en la identificación. Por último, el estado neutral y la tristeza corresponde a un 79.7 % y 90.3 % de clasificaciones acertadas, en tanto que, el 20.3% y 9.7% son diagnósticos erróneos respectivamente.

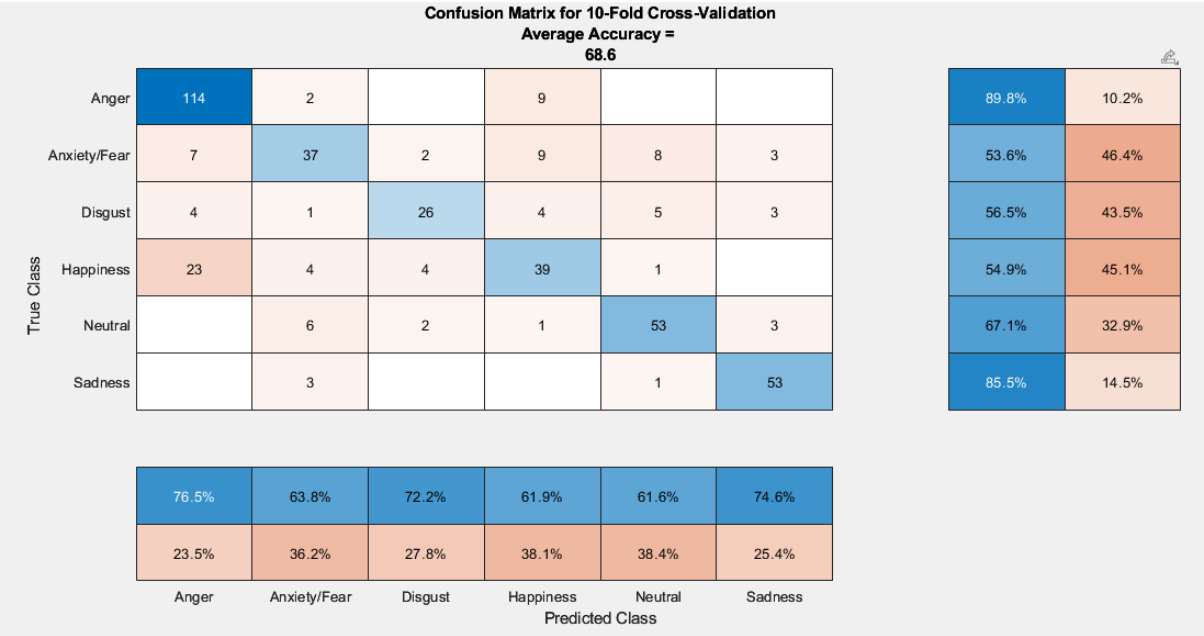
Asimismo se tiene los resultados para los valores predichos en donde se tiene una efectividad de 77.6% y 22.2% de error para la ira, seguidamente se tiene un 74% de efectividad para el miedo con un error de 26%, para el disgusto se tiene una eficiencia del 84.4% con un error del 18.6%, para la felicidad se tiene un acierto del 60% con un error del 40%, asimismo se tiene la eficiencia del estado neutral es de 69.2% y error de 30.8%, finalmente se tiene 73.7% para la tristeza con un error de 26.3%.

### 4.1.2. Modelo 2

Para el segundo modelo planteado se obtiene una matriz de confusión con un resultado de accuracy más bajo que el modelo 1, con un 68.6% la diferencia que se tiene no es muy alta, pero si significativa en este tipo de sistemas con redes neuronales, por lo que se establece el primer modelo para la realización de pruebas de funcionamiento. Los porcentajes de precisión y de error de cada emoción se muestran en la Figura 64.

**Figura 64.**

*Matriz de confusión del Modelo 2*



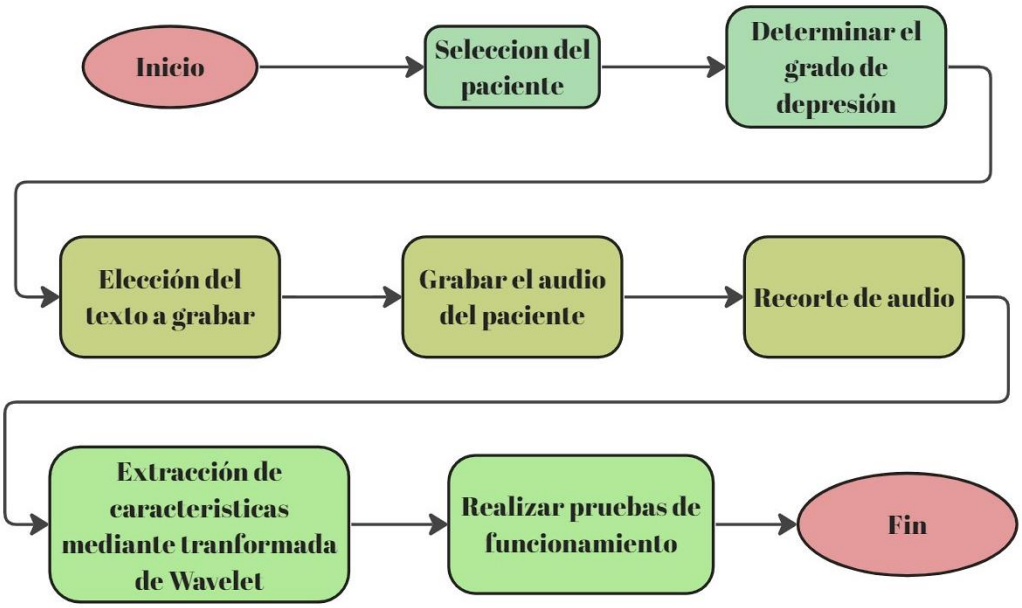

*Fuente:* Elaborado por el autor

## 4.2. Fase de pruebas del sistema

El consentimiento del paciente, el análisis del audio, la extracción de características y la categorización de las emociones forman parte del proceso de las pruebas del sistema. En este sentido se sigue un procedimiento lógico de 7 pasos que se aprecian en la Figura 65.

**Figura 65.**

*Procedimiento para pruebas de funcionamiento del sistema de reconocimiento de emociones*




*Fuente:* Elaborado por el autor

### 4.2.1. Selección del paciente

El muestreo utilizado para la presente investigación es aleatorio o simple, pues permitió escoger sujetos al azar en un conjunto de personas, en este sentido, se evalúa la eficacia del sistema de detección de emociones del habla presentado en este estudio. Para luego realizar un análisis exhaustivo de los resultados de las 6 personas elegidos que se presentan en la Tabla 12.

**Tabla 12**.

*Personas seleccionadas para las pruebas de funcionamiento*

| Nro. | Individuo | Edad |
|---|---|---|
| 1 | Individuo 1 | 18-25 |
| 2 | Individuo 2 | 18-25 |
| 3 | Individuo 3 | 18-25 |
| 4 | Individuo 4 | 18-25 |
| 5 | Individuo 5 | 18-25 |
| 6 | Individuo 6 | 18-25 |

*Fuente:* Elaborado por el autor

### 4.2.2. Determinar el grado de depresión

Se utiliza la prueba de Beck para determinar la patología relacionada con la depresión moderada y grave. Este consta de 21 ítems que abordan síntomas cognitivos, afectivos, motivacionales y fisiológicos de la depresión, que pueden ser utilizados en adultos y adolescentes mayores de 13 años (Ciharova et al., 2020). El test de Beck empleado para esta investigación se adjunta como anexos.

Los valores que se toman en cuenta para detectar el grado de depresión en los pacientes es el que se muestra en la Tabla 13. Se aprecia que si el paciente obtiene un puntaje entre 0 y 13 el grado de depresión que sufre es mínimo por lo tanto el paciente no está sufriendo la patología, asimismo, el puntaje entre 14 y 19 marca una depresión leve que no es preocupante, desde el puntaje 20 a 28 que significa que el paciente tiene una depresión moderada y se debe tener en consideración acudir con el profesional, por último, con el puntaje de 29 a 63 la depresión es grave por lo tanto se debe seguir la terapia recomendada por el especialista.

**Tabla 13**

*Parámetros de evaluación del test de Beck*

| Nro. | Puntaje | Grado de Depresión |
|---|---|---|
| 1 | 0 - 13 | Mínima depresión |
| 2 | 14 - 19 | Depresión Leve |
| 3 | 20 - 28 | Depresión Moderada |
| 4 | 29 - 63 | Depresión Grave |

*Nota.* Tomado de (Ciharova et al., 2020)

### 4.2.3. Elección del texto a grabar

En cuanto a la elección del texto a ser leído por el paciente, se considera un párrafo que contiene una variedad de palabras que despierten varias emociones, ver Figura 66. En este sentido, se trata de que el paciente al momento de leer el texto en mención despierte una emoción y de esta manera poder captarla en audio para su posterior análisis.

**Figura 66.**

*Texto base para lectura de pacientes.*

Nos pasamos toda la vida soñando con deseos incumplidos, recordando cicatrices, construyendo artificial y mentirosamente lo que pudimos haber sido. cada vez somos menos verdaderos, más hipócritas; cada vez tenemos más vergüenza de nuestra verdad

*Fuente:* Elaborado por el autor

### 4.2.4. Grabación de audio del paciente

Para el proceso de grabado del audio se emplea el micrófono HMKCH Lavalier para teléfono celular (USB C), cuenta con dispersión de ruido y graba a una frecuencia de 44kHz, es muy útil ya que al momento de la grabación el paciente debe estar en un lugar aislado, mirar Figura 67.

**Figura 67.**

*Micrófono usado para la grabación.*

*Nota.* Tomado de (Tiendamia, 2023)

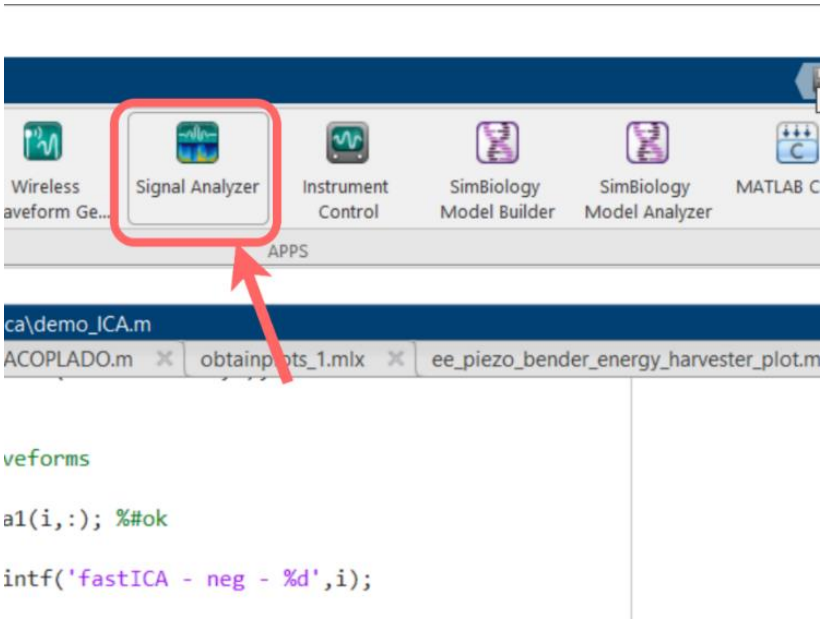
Además, con el fin de que el audio no se vea afectado por ruidos externos al orador, este se graba en un entorno aislado. De esta forma se evita interferencias molestas, mientras que se crea un ambiente propicio para que el paciente se relaje y se produzcan las emociones deseadas. Finalmente, la grabación está realizada por la aplicación de “Grabadora” que viene incluida en el celular y es óptima para el tipo de grabación que se realiza.

### 4.2.5. Recorte de audio grabado

El proceso de fraccionamiento del audio grabado es necesario debido a la duración de cada uno de estos. Dado que, cada paciente habla a un ritmo de habla diferente, se tiene que extraer una parte representativa para ejecutar el análisis. El proceso de fraccionamiento se lleva a cabo con la aplicación “Signal Analyser” de Matlab, observar Figura 68.

**Figura 68.**

*Aplicación Signal Analyser de Matlab*

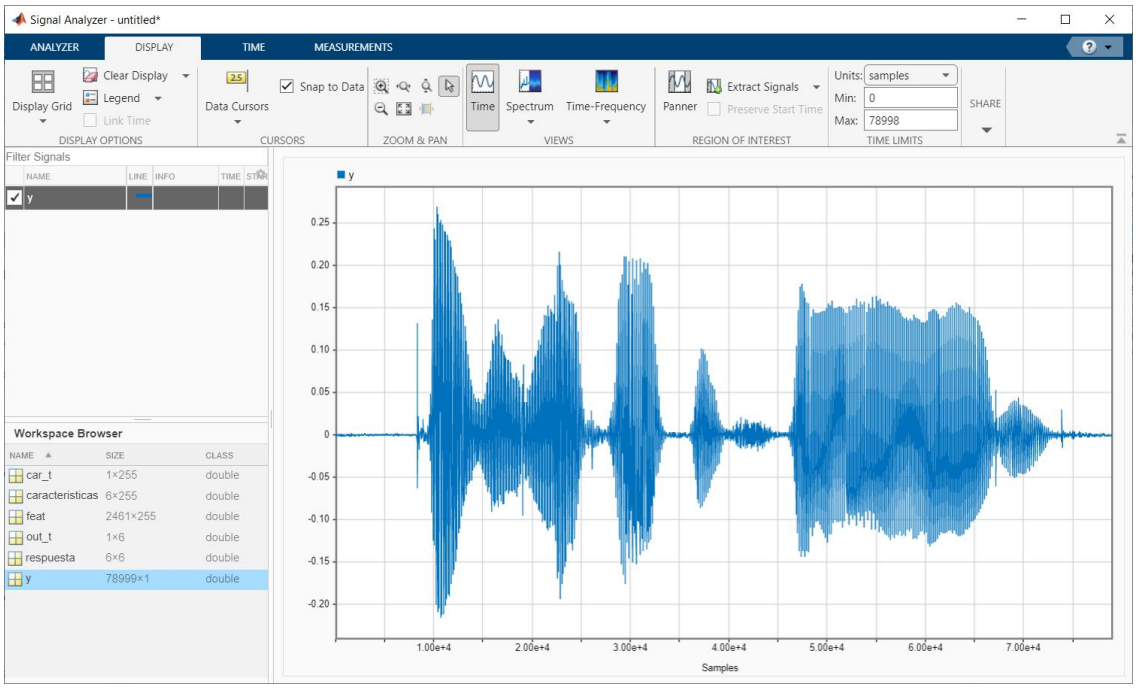



*Fuente:* Elaborado por el autor

Una vez abierto se importa la señal de audio correspondiente, mirar Figura 69, para luego, colocar los cursores en la sección que se desea conservar como se observa en la Figura 70 y, finalmente, extraer la nueva onda que se aprecia en la Figura 71. Una vez extraída se guarda para su posterior análisis.

**Figura 69.**

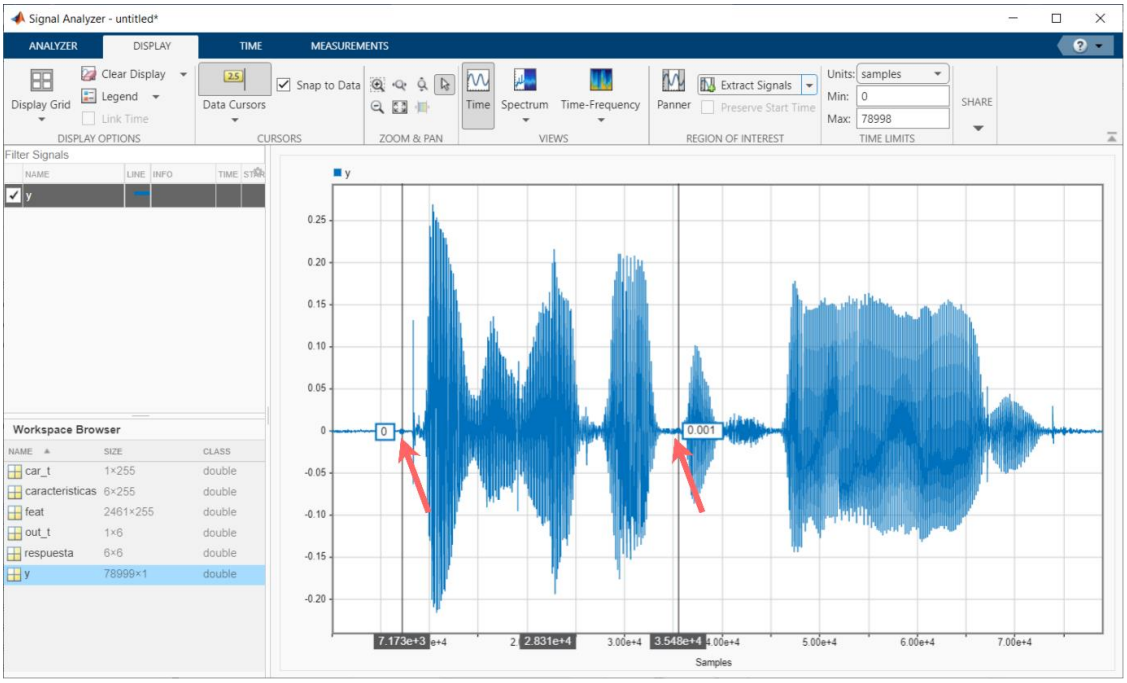
*Audio cargado a la aplicación Signal Analyser de Matlab*




*Fuente:* Elaborado por el autor

**Figura 70.**

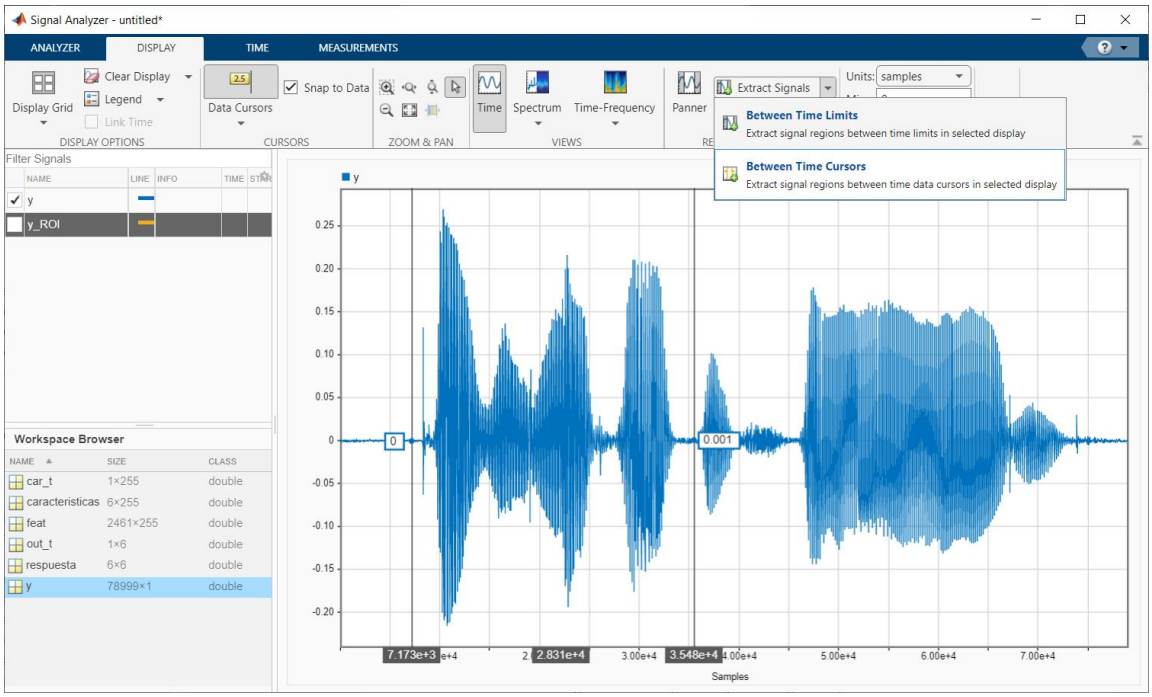
*Selección de intervalo de análisis en Aplicación Signal Analyser.*




*Fuente:* Elaborado por el autor

**Figura 71.**

*Extracción de onda para análisis en Aplicación Signal Analyser.*



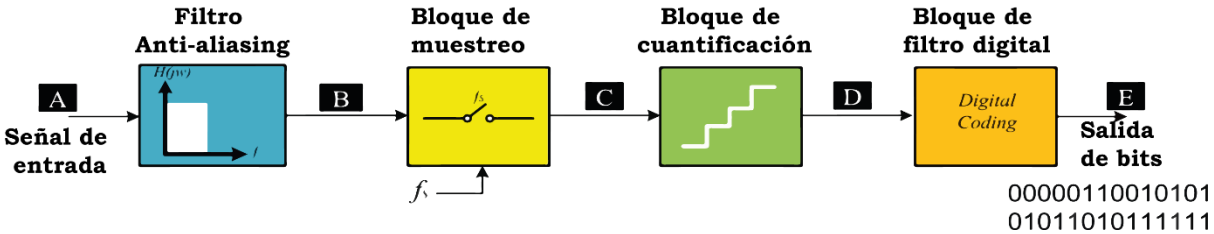

*Fuente:* Elaborado por el autor

### 4.2.6. Proceso de conversión de analógico a digital

Para procesar los datos digitales, la señal analógica debe convertirse a la forma digital mediante convertidores de analógico a digital (ADC). Todos los ADC actuales obedecen las mismas reglas para producir sus bits de salida. En este contexto, existen cuatro bloques principales que se ilustran en la Figura 72 y pueden expresar todo el rendimiento de los ADC.

**Figura 72.**

*Procedimiento ADC*




*Nota.* Adaptado de (Sandoval, 2019)

**Filtro anti-aliasing:** el primer bloque al que se enfrenta la señal de entrada analógica (A) es el filtro de paso bajo. Se emplea este filtro porque se debe eliminar todos los componentes de

frecuencia (armónicos) de la señal analógica que excedan la frecuencia de la tasa de Nyquist. El problema ocurre cuando el espectro de la señal de entrada se superpone con el espectro sin filtrar de las otras señales (armónicos). Para evitar un error de aliasing, el filtro debe eliminar todas las frecuencias analógicas por encima de la frecuencia mínima en el espectro de muestreo.

**Bloque de muestreo:** el segundo es el bloque de muestreo que muestrea la señal de entrada (A) con el intervalo de tiempo de Ts (Ts=1/fs) y genera una señal de tiempo discreta (C).

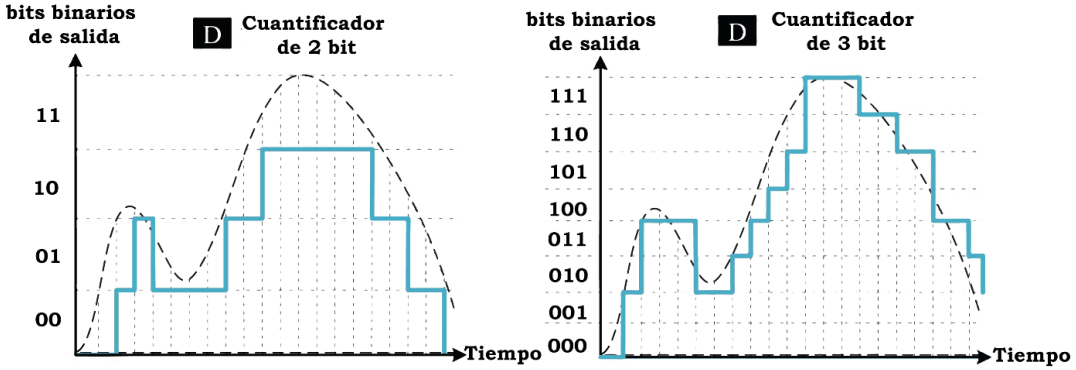
Bloque de cuantificación; (Sección D): Este bloque juega un papel crucial en la estructura de los ADC.

**Bloque de cuantificación:** este bloque juega un papel crucial en la estructura de los ADC.

Para ilustrar más sobre este proceso de cuantificación, se explica el cuantificador de 2 y 3 bits como se presenta en la Figura 73. En este contexto, en el cuantificador de 2 bits y 3 bits se tiene 4 niveles (2^2) y 8 niveles (2^3), respectivamente. Cada intervalo de muestra se numera a lo largo del eje horizontal. Los datos muestreados se conservan durante todo el período de muestreo.

**Figura 73.**

*Proceso de cuantificación*




*Nota.* Adaptado de (Sandoval, 2019)

**Bloque de codificación digital**: este bloque se utiliza en algunos de los ADC que necesitan más proceso sobre el bit de salida como el filtro de decimación en los ADC Sigma-delta.

Para los audios que se obtuvo en la fase de pruebas del sistema se tiene las características que se muestran en la Figura 74.

En donde podemos apreciar la frecuencia de muestro que está determinada en 48kHz, también se tiene un total de muestras de 78999, asimismo la duración del audio que es de 1.6s y los bits por muestra que se obtiene un total de 16 bits.

**Figura 74.**

*Características de los audios obtenidos*

```
info = struct with fields:
             Filename: 'C:\Users\PERSONAL\Desktop\8VO SEMESTRE\TITULACIÓN II\Tesis\Pruebas\Redimensionado\pc_1.wav
    CompressionMethod: 'Uncompressed'
          NumChannels: 1
           SampleRate: 48000
         TotalSamples: 78999
             Duration: 1.6458
                Title: []
              Comment: []
               Artist: []
        BitsPerSample: 16
```

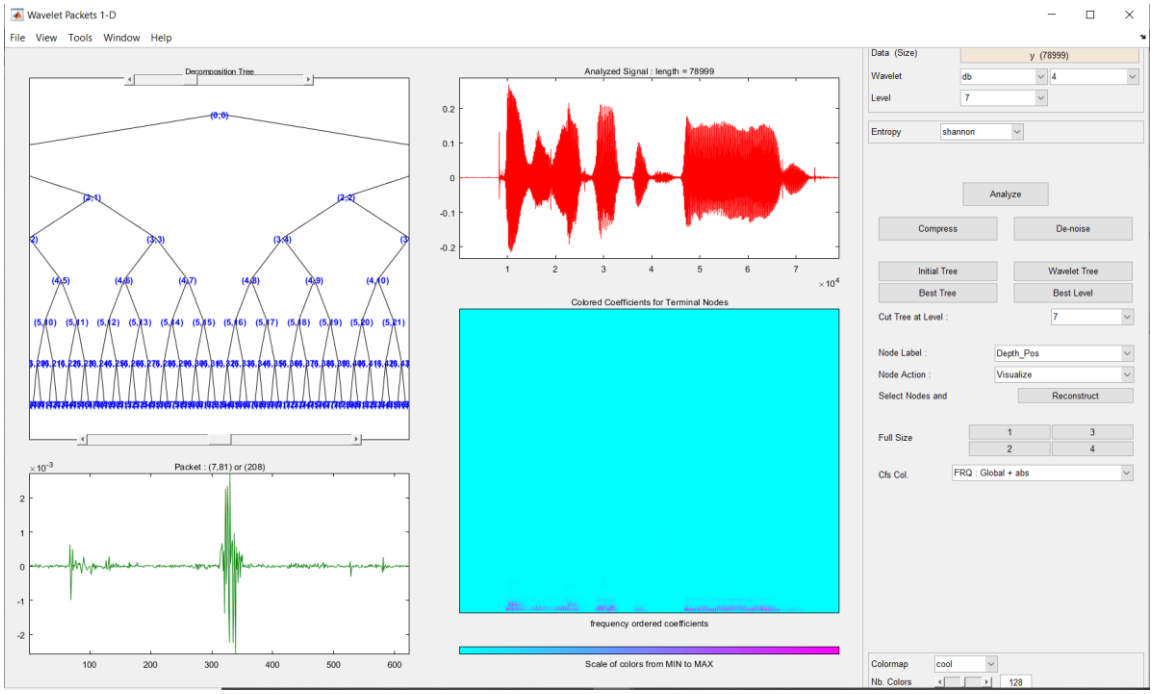
*Fuente:* Elaborado por el autor

### 4.2.7. Extracción de características

El procesado de obtención de parámetros característicos se basa en la transformada de wavelet según su árbol de descomposición. En este caso en particular se usa la familia DB grado cuatro con 7 niveles de composición como señala en la Figura 75.

**Figura 75.**

*Descomposición de wavelet*

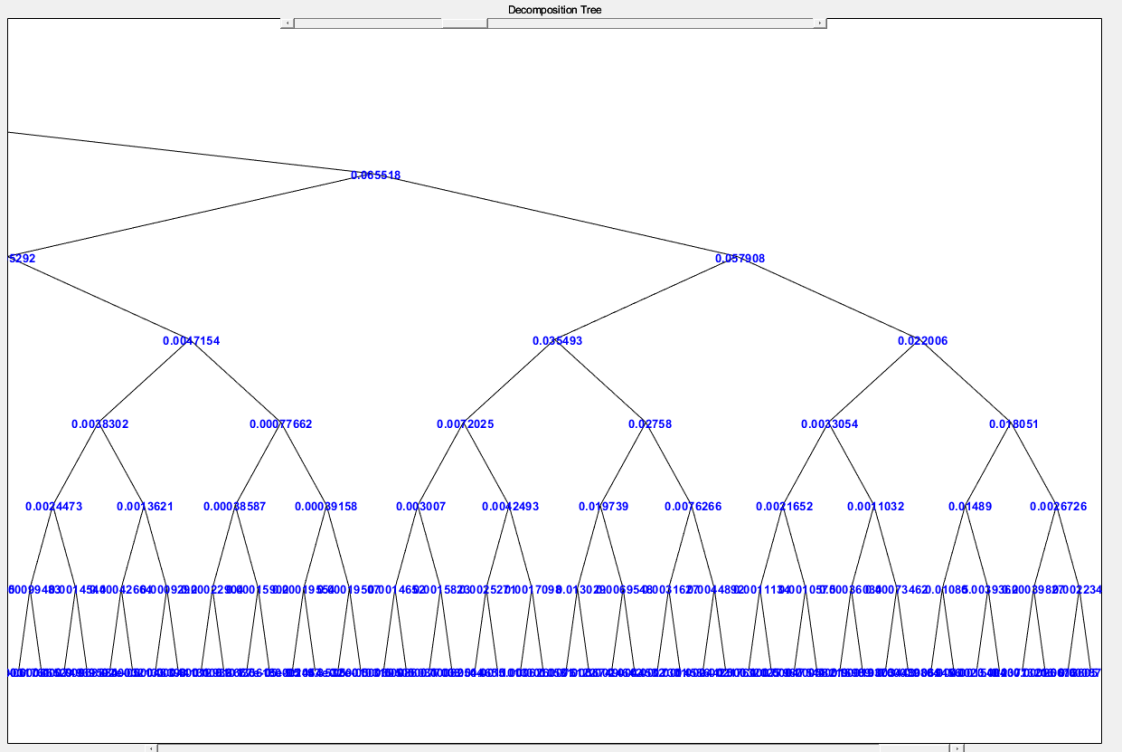



*Fuente:* Elaborado por el autor

Por otro lado, se calcula la energía en cada nodo de cada nivel de descomposición, puesto que estas serán las características del audio analizado, mirar Figura 76.

**Figura 76.**

*Energía nodal de cada nivel de descomposición wavelet*

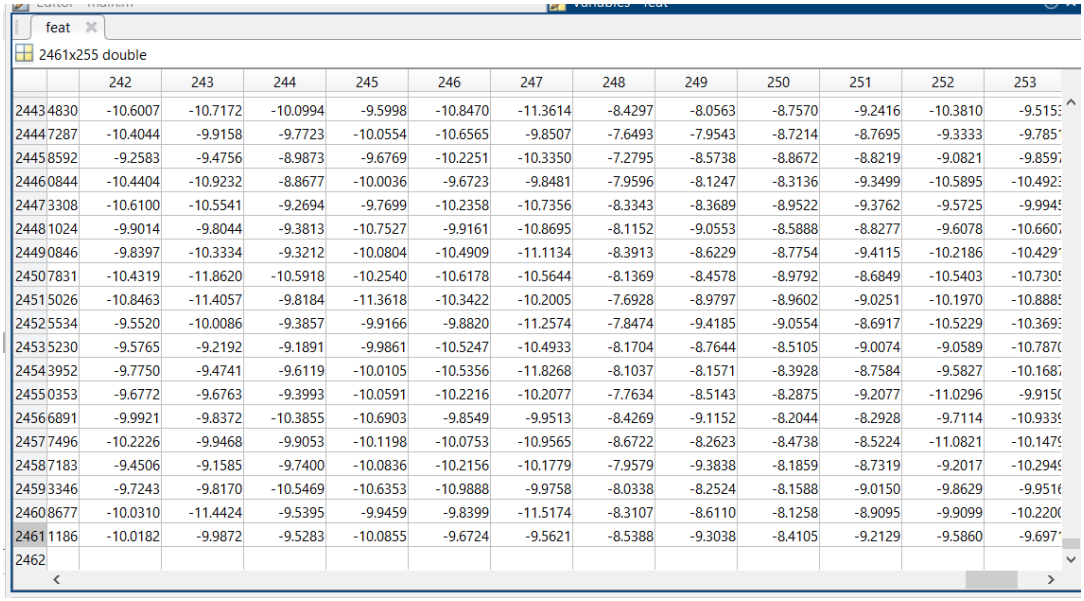



*Fuente:* Elaborado por el autor

Por último, los valores de energía cuantificados se almacenan en el espacio de trabajo de Matlab a forma de matriz, en la Figura 77 se observa la matriz con las 255 señales extraídas en donde se tiene 2461 elementos en cada señal.

**Figura 77.**

*Matriz de descomposición wavelet nivel 7 perteneciente a un audio*



feat

2461x255 double

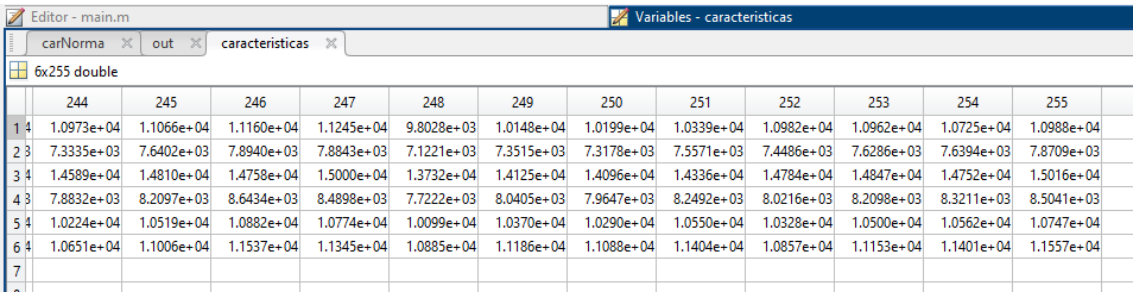
| | | 242 | 243 | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 24434830 | | -10.6007 | -10.7172 | -10.0994 | -9.5998 | -10.8470 | -11.3614 | -8.4297 | -8.0563 | -8.7570 | -9.2416 | -10.3810 | -9.515 |
| 24447287 | | -10.4044 | -9.9158 | -9.7723 | -10.0554 | -10.6565 | -9.8507 | -7.6493 | -7.9543 | -8.7214 | -8.7695 | -9.3333 | -9.785 |
| 24458592 | | -9.2583 | -9.4756 | -8.9873 | -9.6769 | -10.2251 | -10.3350 | -7.2795 | -8.5738 | -8.8672 | -8.8219 | -9.0821 | -9.859 |
| 24460844 | | -10.4404 | -10.9232 | -8.8677 | -10.0036 | -9.6723 | -9.8481 | -7.9596 | -8.1247 | -8.3136 | -9.3499 | -10.5895 | -10.492 |
| 24473308 | | -10.6100 | -10.5541 | -9.2694 | -9.7699 | -10.2358 | -10.7356 | -8.3343 | -8.3689 | -8.9522 | -9.3762 | -9.5725 | -9.994 |
| 24481024 | | -9.9014 | -9.8044 | -9.3813 | -10.7527 | -9.9161 | -10.8695 | -8.1152 | -9.0553 | -8.5888 | -8.8277 | -9.6078 | -10.660 |
| 24490846 | | -9.8397 | -10.3334 | -9.3212 | -10.0804 | -10.4909 | -11.1134 | -8.3913 | -8.6229 | -8.7754 | -9.4115 | -10.2186 | -10.429 |
| 24507831 | | -10.4319 | -11.8620 | -10.5918 | -10.2540 | -10.6178 | -10.5644 | -8.1369 | -8.4578 | -8.9792 | -8.6849 | -10.5403 | -10.730 |
| 24515026 | | -10.8463 | -11.4057 | -9.8184 | -11.3618 | -10.3422 | -10.2005 | -7.6928 | -8.9797 | -8.9602 | -9.0251 | -10.1970 | -10.888 |
| 24525534 | | -9.5520 | -10.0086 | -9.3857 | -9.9166 | -9.8820 | -11.2574 | -7.8474 | -9.4185 | -9.0554 | -8.6917 | -10.5229 | -10.369 |
| 24535230 | | -9.5765 | -9.2192 | -9.1891 | -9.9861 | -10.5247 | -10.4933 | -8.1704 | -8.7644 | -8.5105 | -9.0074 | -9.0589 | -10.787 |
| 24543952 | | -9.7750 | -9.4741 | -9.6119 | -10.0105 | -10.5356 | -11.8268 | -8.1037 | -8.1571 | -8.3928 | -8.7584 | -9.5827 | -10.168 |
| 24550353 | | -9.6772 | -9.6763 | -9.3993 | -10.0591 | -10.2216 | -10.2077 | -7.7634 | -8.5143 | -8.2875 | -9.2077 | -11.0296 | -9.915 |
| 24566891 | | -9.9921 | -9.8372 | -10.3855 | -10.6903 | -9.8549 | -9.9513 | -8.4269 | -9.1152 | -8.2044 | -8.2928 | -9.7114 | -10.933 |
| 24577496 | | -10.2226 | -9.9468 | -9.9053 | -10.1198 | -10.0753 | -10.9565 | -8.6722 | -8.2623 | -8.4738 | -8.5224 | -11.0821 | -10.147 |
| 24587183 | | -9.4506 | -9.1585 | -9.7400 | -10.0836 | -10.2156 | -10.1779 | -7.9579 | -9.3838 | -8.1859 | -8.7319 | -9.2017 | -10.294 |
| 24593346 | | -9.7243 | -9.8170 | -10.5469 | -10.6353 | -10.9888 | -9.9758 | -8.0338 | -8.2524 | -8.1588 | -9.0150 | -9.8629 | -9.951 |
| 24608677 | | -10.0310 | -11.4424 | -9.5395 | -9.9459 | -9.8399 | -11.5174 | -8.3107 | -8.6110 | -8.1258 | -8.9095 | -9.9099 | -10.220 |
| 24611186 | | -10.0182 | -9.9872 | -9.5283 | -10.0855 | -9.6724 | -9.5621 | -8.5388 | -9.3038 | -8.4105 | -9.2129 | -9.5860 | -9.697 |
| 2462 | | | | | | | | | | | | | |

*Fuente:* Elaborado por el autor

Finalmente, se realiza el cálculo de energía para cada una de las señales resultantes por lo que se obtiene un vector de 255 características, en la Figura 78 se obtiene las características de los audios de los 6 pacientes de los cuales se realizaron las pruebas. Luego a estos datos se le realiza la respectiva normalización que se explicó en el capítulo anterior para poder entrar a la red neuronal.

**Figura 78.**

*Matriz de características de los audios de los 6 pacientes*



Editor - main.m

Variables - caracteristicas

carNorma | out | caracteristicas

6x255 double

| | | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 1.0973e+04 | 1.1066e+04 | 1.1160e+04 | 1.1245e+04 | 9.8028e+03 | 1.0148e+04 | 1.0199e+04 | 1.0339e+04 | 1.0982e+04 | 1.0962e+04 | 1.0725e+04 | 1.0988e+04 |
| 2 | 3 | 7.3335e+03 | 7.6402e+03 | 7.8940e+03 | 7.8843e+03 | 7.1221e+03 | 7.3515e+03 | 7.3178e+03 | 7.5571e+03 | 7.4486e+03 | 7.6286e+03 | 7.6394e+03 | 7.8709e+03 |
| 3 | 4 | 1.4589e+04 | 1.4810e+04 | 1.4758e+04 | 1.5000e+04 | 1.3732e+04 | 1.4125e+04 | 1.4096e+04 | 1.4336e+04 | 1.4784e+04 | 1.4847e+04 | 1.4752e+04 | 1.5016e+04 |
| 4 | 3 | 7.8832e+03 | 8.2097e+03 | 8.6434e+03 | 8.4898e+03 | 7.7222e+03 | 8.0405e+03 | 7.9647e+03 | 8.2492e+03 | 8.0216e+03 | 8.2098e+03 | 8.3211e+03 | 8.5041e+03 |
| 5 | 4 | 1.0224e+04 | 1.0519e+04 | 1.0882e+04 | 1.0774e+04 | 1.0099e+04 | 1.0370e+04 | 1.0290e+04 | 1.0550e+04 | 1.0328e+04 | 1.0500e+04 | 1.0562e+04 | 1.0747e+04 |
| 6 | 4 | 1.0651e+04 | 1.1006e+04 | 1.1537e+04 | 1.1345e+04 | 1.0885e+04 | 1.1186e+04 | 1.1088e+04 | 1.1404e+04 | 1.0857e+04 | 1.1153e+04 | 1.1401e+04 | 1.1557e+04 |
| 7 | | | | | | | | | | | | | |

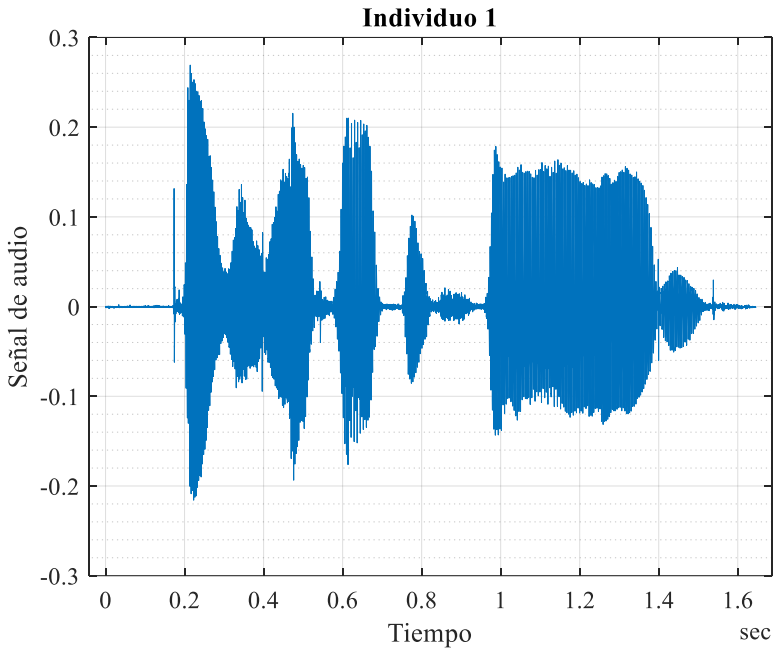
*Fuente:* Elaborado por el autor

### 4.2.8. Pruebas de funcionamiento

El proceso se inicia cuando el paciente acepta participar en el estudio (Anexo 1), indicando que está dispuesto a participar en las pruebas para comprobar la eficacia del sistema de detección de emociones mediante el habla. Posterior a la concesión del permiso por parte del paciente, se procede a ejecutar la prueba de Beck (Anexo 2). A continuación, se capta el audio de la persona, que se analiza y se introduce en la técnica de inteligencia artificial para categorizar la emoción correspondiente.

**Figura 79.**

*Audio original del Paciente 1*



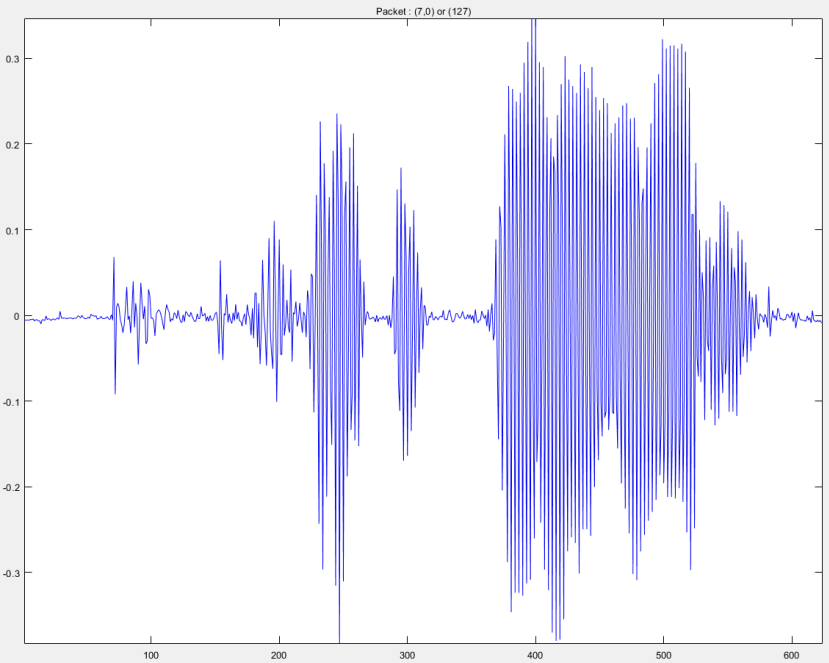

*Fuente:* Elaborado por el autor

Las respuestas obtenidas del test de Beck para el paciente 1 indica que el grado de depresión que presenta es grave. Esto es coherente con una emoción de tristeza y desanimo que se puede apreciar en la Figura 84. En este sentido, se constata que la forma de onda describe pausas, lo que se asocia con un ritmo de habla más lento y un tono medio más bajo. Asimismo, se presentan los coeficientes de aproximación en la Figura 80 y detalle en la Figura 81 respectivamente. En consecuencia, con lo que respecta al coeficiente de aproximación se observa que la amplitud es variable y describe muchas pausas.

Por otro lado, el coeficiente de detalle muestra una onda con una amplitud constante, no obstante, las pausas prevalecen. Cabe recalcar que los coeficientes de aproximación son señales que muestran las bajas frecuencias de la señal original, por lo que para extraer este tipo de señales se usa un filtro pasa bajo.

**Figura 80.**

*Coeficiente de aproximación de paciente 1*

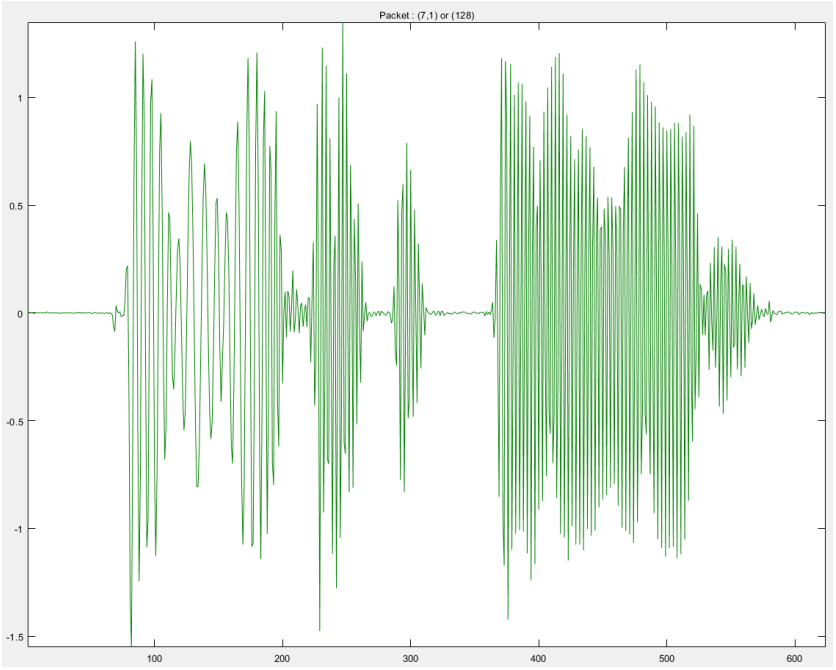



*Fuente:* Elaborado por el autor

Por el contrario, los coeficientes de detalles muestran las altas frecuencias que componen la señal de la que se extrae, usando así un filtro pasa alto para realizar este proceso de extracción.

**Figura 81.**

*Coeficiente de detalle de paciente 1*

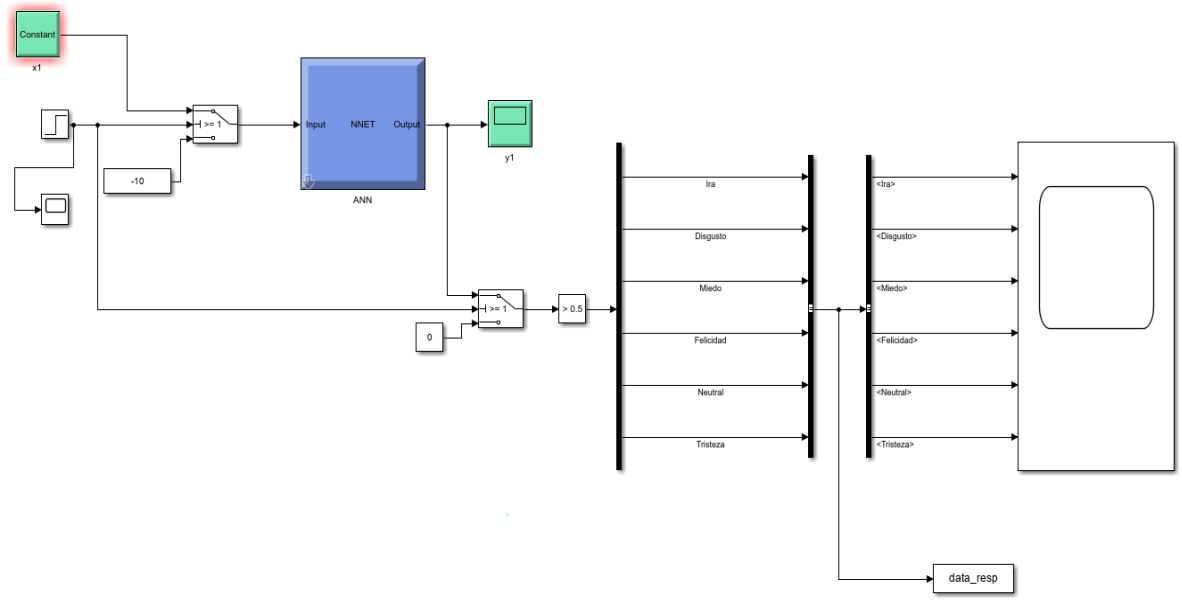



*Fuente:* Elaborado por el autor

Para la prueba de funcionamiento se emplea un diagrama que consta de varios bloques. Entre ellos está la constante (x1), la red neuronal entrenada, el demultiplexor, el creador de arreglos y el scope para la visualización, el diagrama se mira en la Figura82.

**Figura 82.**

*Diagrama de pruebas del sistema de reconocimiento de emociones*

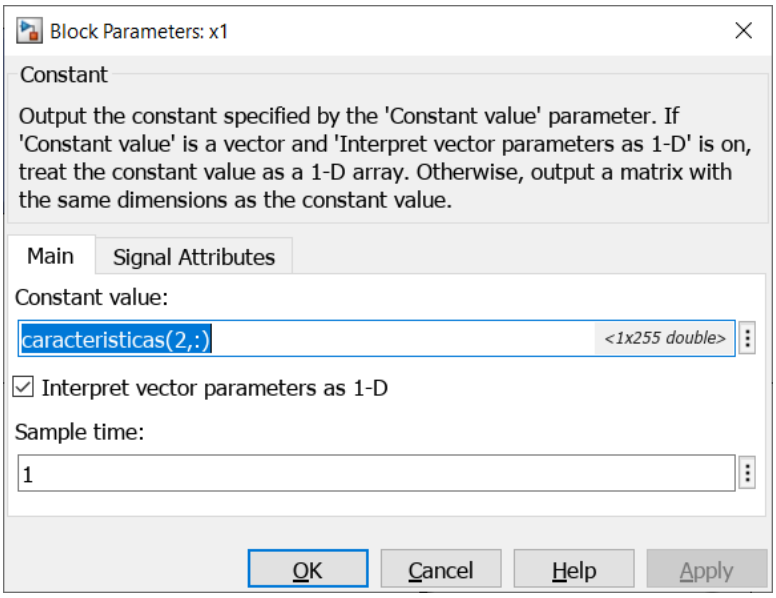



*Fuente:* Elaborado por el autor

Los datos se ingresan en el bloque “constant” tal como se observa en la Figura 83, en donde se debe colocar las 255 características del audio que se desea predecir la emoción.

**Figura 83.**

*Datos del paciente 1 en el sistema de reconocimiento de emociones*

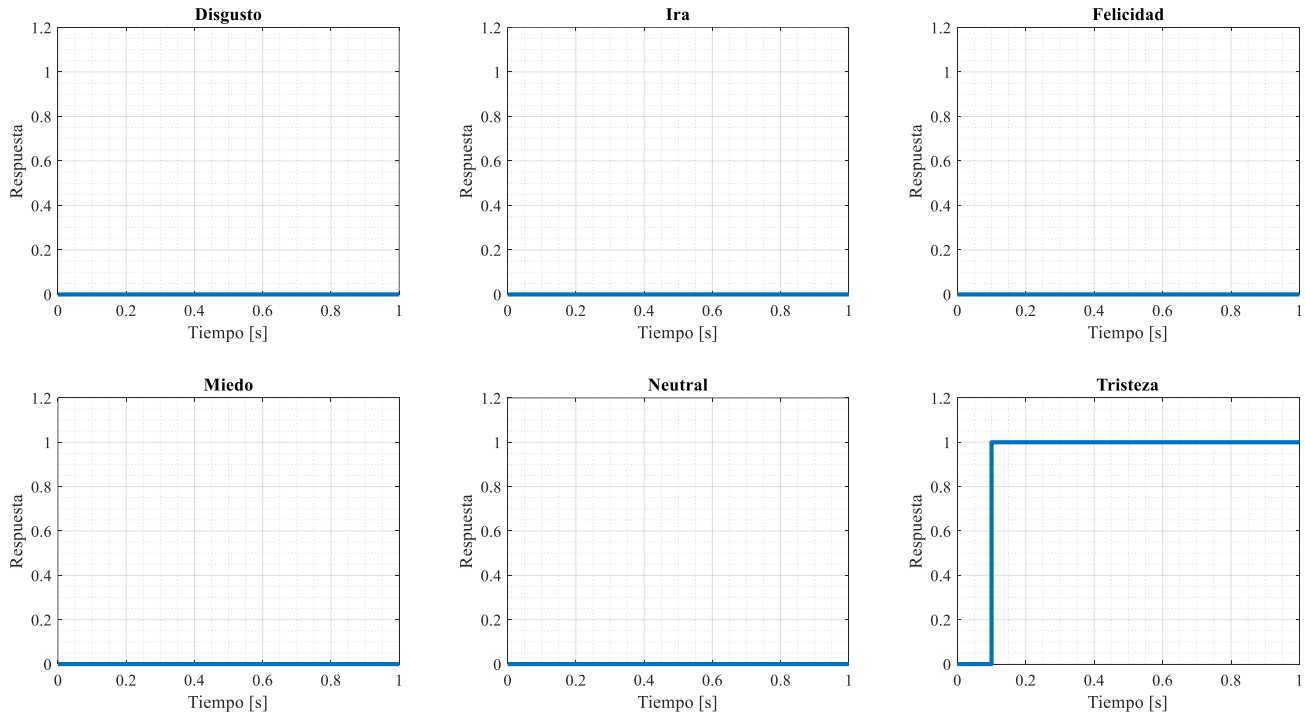



*Fuente:* Elaborado por el autor

La respuesta de la red neuronal indica que el audio ingresado corresponde a una emoción de tristeza como se muestra en la Figura 84. Este resultado es coherente con una depresión grave como se constató con anterioridad con el test de Beck. El diagrama se lo representa conforme se realizó el etiquetado de los audios, ya que al momento de realizar una predicción la red neuronal arroja valores en sus 6 capas de salida, estos valores van del 0 al 1 y el que más se aproxima a 1 tiende a ser el resultado esperado, por lo que en el diagrama de simulink se usa un sumador para que estos valores los redondee a 1 o a 0 y así obtener la gráfica de resultado.

**Figura 84.**

*Respuesta del sistema de reconocimiento de emociones para el Individuo 1*




*Fuente:* Elaborado por el autor

Por último, en la Tabla 14 se presenta el resultado de las emociones detectadas en cada paciente, en donde se detalla la edad, la emoción que detecto el sistema, el grado de depresión que se obtuvo con el test realizado y el puntaje que se tiene con este test.

**Tabla 14**

*Resumen de los resultados obtenidos en la red neuronal*

| Nro. | Individuo | Género | Edad | Emoción | Patología | Puntaje del test |
|---|---|---|---|---|---|---|
| 1 | Individuo 1 | Femenino | 18-25 | Tristeza | Depresión Grave | 36 |
| 2 | Individuo 2 | Masculino | 18-25 | Miedo | Depresión leve | 13 |
| 3 | Individuo 3 | Masculino | 18-25 | Miedo | Depresión leve | 16 |
| 4 | Individuo 4 | Femenino | 18-25 | Miedo | Depresión leve | 18 |
| 5 | Individuo 5 | Masculino | 18-25 | Tristeza | Depresión Grave | 37 |
| 6 | Individuo 6 | Femenino | 18-25 | Tristeza | Depresión Grave | 53 |

*Fuente:* Elaborado por el autor

# CONCLUSIONES

El sistema de reconocimiento de emociones a través de la voz es viable, ya que se estableció una relación entre las emociones detectadas por el sistema y el test aplicado a los pacientes, logrando así poder reconocer el nivel de depresión que tienen los pacientes. En este sentido se puede afirmar que aproximadamente un 66.6% de las mujeres de la muestra se estima que padecen depresión grave, mientras que en los hombres se tiene alrededor del 33.3% se estima que padecen dicha patología. Adicionalmente se corroboró la factibilidad de técnicas de aprendizaje profundo, en este caso la de redes neuronales.

Con la revisión bibliográfica realizada se logró contrastar las ideas de diferentes autores para poder realizar un análisis de los temas fundamentales que componen la investigación, principalmente temas como la representación del sonido en el habla, la extracción de características, las redes neuronales y la relación de estas con el reconocimiento de emociones, finalizando con el diagnóstico de la depresión.

La arquitectura KDD determino el diseño del sistema, empezando por la obtención de la base de datos, el procesamiento de los audios, la extracción de características y el entrenamiento de la red. En el proceso de entrenamiento de la red neuronal, se establecieron dos modelos, los cuales fueron sometidos a un análisis para determinar el óptimo para la investigación, en base a esto, se obtuvieron los parámetros que afectan significativamente el entrenamiento de la red neuronal que son: algoritmo de resolución (descenso del gradiente), número de épocas (3), tamaño mínimo de lote (217) y taza de aprendizaje (0.005).

La red neuronal presenta una efectividad del 73.5%, en cuanto a la efectividad y el error en cada emoción se obtuvo los siguientes resultados, la ira 82.7%.30 % con una tasa de error de 17.3%. Asimismo, el disgusto presenta un 76.1 % de eficacia, junto con un 23.9% de clasificaciones erróneas. Del mismo modo, el miedo representa un 53.6 % de detecciones correctas, mientras que, el 46.4 % se asocia con una categoría errónea. De manera similar, la

felicidad posee una eficiencia de 50.7 %, con un 49.3 % de errores en la identificación. Por último, el estado neutral y la tristeza corresponde a un 79.7 % y 90.3 % de clasificaciones acertadas, en tanto que, el 20.3% y 9.7% son diagnósticos erróneos respectivamente

La efectividad de la red neuronal depende en gran medida de la calidad de la extracción de las características de los audios de emociones, así como de la parametrización empleada para el entrenamiento de la red. Esto solo se puede determinar mediante la experimentación y en base a trabajos similares que se encuentran en la comunidad científica.

# RECOMENDACIONES

Se recomienda embeber el programa en hardware para su posible implementación en las instituciones que se dediquen a atender pacientes con depresión, puesto que se comprueba que es una herramienta útil, que permite alertar y diagnosticar patologías asociadas con la depresión grave y leve por medio de las emociones.

Realizar un estudio comparativo de los resultados obtenidos en la presente investigación con diferentes técnicas para la detección de emociones mediante la voz. Con el fin de evaluar la efectividad a comparación con otras técnicas e identificar posibles mejoras que se puedan implementar en la solución propuesta.

Realizar un estudio con otro tipo de datos que representen las señales de audio de entrada, debido a que en esta investigación se hizo uso de la extracción de características de las formas de onda por medio de la transformada de Wavelet, tomando así datos numéricos los cuales representaban la energía característica de cada señal procedente de la transformada de Wavelet multinivel.

Crear un data set con audios de pacientes reales que estén cruzando la patología de depresión, esto para tener un conjunto de datos especifico con la patología y así diseñar un sistema más confiable en la detección de la patología para finalmente implementarlo en consultorios o centros dedicados a esta enfermedad, ya que los data sets disponibles son muy limitados.

# Bibliografía

Aggarwal, A., Srivastava, A., Agarwal, A., Chahal, N., Singh, D., Alnuaim, A. A., Alhadlaq, A., & Lee, H. (2022). Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. *Sensors*, *22*(6), 2378. https://doi.org/10.3390/s22062378

Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, *43*(2), 155–177. https://doi.org/10.1007/s10462-012-9368-5

Aouani, H., & Ayed, Y. Ben. (2020). Speech Emotion Recognition with deep learning. *Procedia Computer Science*, *176*, 251–260. https://doi.org/10.1016/j.procs.2020.08.027

Bailey, A., & Plumbley, M. D. (2021). Gender Bias in Depression Detection Using Audio Features. *European Signal Processing Conference*, *2021-Augus*, 596–600. https://doi.org/10.23919/EUSIPCO54536.2021.9615933

Brown, R. D. (2013). LNAI 8082 - Text, Speech, and Dialogue. In *Text, Speech and Dialogue. Lecture Notes in Computer Science* (Issue September).

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *Interspeech 2005*, *September*, 1517–1520. https://doi.org/10.21437/Interspeech.2005-446

Byun, S., & Lee, S. (2021). A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms. *Applied Sciences*, *11*(4), 1890. https://doi.org/10.3390/app11041890

Calin, O. (2020). *Deep Learning Architectures* (1ª ed). Springer. https://doi.org/10.1007/978-3-030-36721-3_5

Caschera, M. C., Grifoni, P., & Ferri, F. (2022). Emotion Classification from Speech and Text

in Videos Using a Multimodal Approach. *Multimodal Technologies and Interaction*, *6*(4), 28. https://doi.org/10.3390/mti6040028

Chen, X., Levitan, S. I., Levine, M., Mandic, M., & Hirschberg, J. (2020). Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies. *Transactions of the Association for Computational Linguistics*, *8*, 199–214. https://doi.org/10.1162/tacl_a_00311

Chowdhary, K. R. (2020). Fundamentals of artificial intelligence. In *Fundamentals of Artificial Intelligence* (1ª ed). Springer. https://doi.org/10.1007/978-81-322-3972-7

Ciharova, M., Cígler, H., Dostálová, V., Šivicová, G., & Bezdicek, O. (2020). Beck depression inventory, second edition, Czech version: demographic correlates, factor structure and comparison with foreign data. *International Journal of Psychiatry in Clinical Practice*, *24*(4), 371–379. https://doi.org/10.1080/13651501.2020.1775854

Costantini, G., Parada-Cabaleiro, E., Casali, D., & Cesarini, V. (2022). The Emotion Probe: On the Universality of Cross-Linguistic and Cross-Gender Speech Emotion Recognition via Machine Learning. *Sensors*, *22*(7). https://doi.org/10.3390/s22072461

Costantini, L., Pasquarella, C., Odone, A., Colucci, M. E., Costanza, A., Serafini, G., Aguglia, A., Belvederi Murri, M., Brakoulias, V., Amore, M., Ghaemi, S. N., & Amerio, A. (2021). Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): A systematic review. *Journal of Affective Disorders*, *279*, 473–483. https://doi.org/10.1016/j.jad.2020.09.131

Deng, L., & Yu, D. (2013). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, *7*(3–4), 197–387. https://doi.org/10.1561/2000000039

Dileep, A. D., & Sekhar, C. C. (2014). GMM-based intermediate matching kernel for

classification of varying length patterns of long duration speech using support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(8), 1421–1432. https://doi.org/10.1109/TNNLS.2013.2293512

Fernandes, B., & Mannepalli, K. (2021). Speech emotion recognition using deep learning lstm for tamil language. *Pertanika Journal of Science and Technology*, *29*(3), 1915–1936. https://doi.org/10.47836/pjst.29.3.33

Ganapathy, A. (2016). Speech Emotion Recognition Using Deep Learning Techniques. *ABC Journal of Advanced Research*, *5*(2), 113–122. https://doi.org/10.18034/abcjar.v5i2.550

Guo, J. (2022). Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems*, *31*(1), 113–126. https://doi.org/10.1515/jisys-2022-0001

Hansen, L., Zhang, Y. P., Wolf, D., Sechidis, K., Ladegaard, N., & Fusaroli, R. (2022). A generalizable speech emotion recognition model reveals depression and remission. *Acta Psychiatrica Scandinavica*, *145*(2), 186–199. https://doi.org/10.1111/acps.13388

Haton, J. P. (2003). Automatic speech recognition: A Review. *ICEIS 2003 - Proceedings of the 5th International Conference on Enterprise Information Systems*, *1*(9), IS5–IS10. https://doi.org/10.5120/9722-4190

Jahangir, R., Teh, Y. W., Hanif, F., & Mujtaba, G. (2021). Deep learning approaches for speech emotion recognition: state of the art and research challenges. In *Multimedia Tools and Applications* (Vol. 80, Issue 16). Multimedia Tools and Applications. https://doi.org/10.1007/s11042-020-09874-7

Jha, M. (2022). *Smart Intelligent Computing and Applications, Volume 1* (V. Bhateja, S. C. Satapathy, C. M. Travieso-Gonzalez, & T. Adilakshmi (eds.); Vol. 282, Issue Sci).

Springer Nature Singapore. https://doi.org/10.1007/978-981-16-9669-5

Jo, T. (2021). *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning* (1ª ed). Springer.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access*, *7*, 117327–117345. https://doi.org/10.1109/ACCESS.2019.2936124

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2003). The patient health questionnaire-2: Validity of a two-item depression screener. *Medical Care*, *41*(11), 1284–1292. https://doi.org/10.1097/01.MLR.0000093487.78664.3C

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, D., Sun, L., Xu, X., Wang, Z., Zhang, J., & Du, W. (2021). BLSTM and CNN Stacking Architecture for Speech Emotion Recognition. *Neural Processing Letters*, *53*(6), 4097–4115. https://doi.org/10.1007/s11063-021-10581-z

Mahony, N. O., Campbell, S., Carvalho, A., Harapanahalli, S., Velasco-Hernandez, G., Krpalkova, L., Riordan, D., & Walsh, J. (2019). Deep Learning vs. Traditional Computer Vision. *Computer Vision*, *Cv*. https://doi.org/10.1007/978-3-030-17795-9

Maji, B., Swain, M., & Mustaqeem, M. (2022). Advanced Fusion-Based Speech Emotion Recognition System Using a Dual-Attention Mechanism with Conv-Caps and Bi-GRU Features. *Electronics*, *11*(9), 1328. https://doi.org/10.3390/electronics11091328

Manoret, P., Chotipurk, P., Sunpaweravong, S., Jantrachotechatchawan, C., & Kobchai, D. (2021). Automatic detection of depression from stratified od audio data. *Journal of JCS Cardiologists*, *12*(2), 345–350. https://doi.org/10.1253/jjcsc.12.2_345

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 Audio-Visual Emotion Database. *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, *1*, 8–8. https://doi.org/10.1109/ICDEW.2006.145

Mauchand, M., & Pell, M. D. (2021). Emotivity in the Voice: Prosodic, Lexical, and Cultural Appraisal of Complaining Speech. *Frontiers in Psychology*, *11*(January), 1–13. https://doi.org/10.3389/fpsyg.2020.619222

Meng, H., Yan, T., Wei, H., & Ji, X. (2021). Speech emotion recognition using wavelet packet reconstruction with attention-based deep recurrent neutral networks. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, *69*(1), 1–12. https://doi.org/10.24425/bpasts.2020.136300

Miikkulainen, R. (2017). Topology of a Neural Network. In *Encyclopedia of Machine Learning and Data Mining* (pp. 1281–1281). https://doi.org/10.1007/978-1-4899-7687-1_843

Moine, C. Le, Obin, N., & Roebel, A. (2021). Speaker Attentive Speech Emotion Recognition. *Interspeech 2021*, 2866–2870. https://doi.org/10.21437/Interspeech.2021-573

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*, *7*, 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

Niu, Y., Zou, D., Niu, Y., He, Z., & Tan, H. (2017). A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks. *A Breakthrough in Speech Emotion Recognition Using Deep Retinal Convolution Neural Networks*, 1–7.

https://doi.org/https://doi.org/10.48550/arXiv.1707.09917

OMS. (2021). *Depresión*.

Patel, J., & Goyal, R. (2008). Applications of Artificial Neural Networks in Medical Science. *Current Clinical Pharmacology*, *2*(3), 217–226. https://doi.org/10.2174/157488407781668811

Rázuri, J., Sundgren, D., Rahmani, R., Larsson, A., Moran, A., & Bonet, I. (2015). Speech emotion recognition in emotional feedback for Human-Robot Interaction. *International Journal of Advanced Research in Artificial Intelligence*, *4*(2). https://doi.org/10.14569/ijarai.2015.040204

Rintala, J. (2020). Speech Emotion Recognition from Raw Audio using Deep Learning. *Degree Project Computer Science and Engineering*.

Sandoval, V. (2019). *Reconocimiento de emociones por medio de la voz*. Universidad Politécnica de Madrid.

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schuller, B. W. (2018). Speech Emotion Recognition two decades in a Nutshell. *Communications of the ACM*, *61*(5), 90–99.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). *Compute Trends Across Three Eras of Machine Learning*. http://arxiv.org/abs/2202.05924

Singh, A., Kaur, N., Kukreja, V., Kadyan, V., & Kumar, M. (2022). Computational intelligence in processing of speech acoustics : a survey. *Complex & Intelligent Systems*. https://doi.org/10.1007/s40747-022-00665-1

Sisman, B. (2019). *Machine Learning for Limited Data Voice Conversion*. National University of Singapore.

Sri Lalitha, Y., Basha Sk, A. H., & Aditya Nag, M. V. (2021). Neural Network Modelling of Speech Emotion Detection. *E3S Web of Conferences*, *309*, 01139. https://doi.org/10.1051/e3sconf/202130901139

Sultana, S., Rahman, M. S., Selim, M. R., & Iqbal, M. Z. (2021). SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *PLoS ONE*, *16*(4 April), 1–27. https://doi.org/10.1371/journal.pone.0250173

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, *21*(1), 93–120. https://doi.org/10.1007/s10772-018-9491-z

Tătaru, O. S., Vartolomei, M. D., Rassweiler, J. J., Virgil, O., Lucarelli, G., Porpiglia, F., Amparore, D., Manfredi, M., Carrieri, G., Falagario, U., Terracciano, D., de Cobelli, O., Busetto, G. M., Del Giudice, F., & Ferro, M. (2021). Artificial intelligence and machine learning in prostate cancer patient management—current trends and future perspectives. In *Diagnostics* (Vol. 11, Issue 2, pp. 1–20). https://doi.org/10.3390/diagnostics11020354

Toshinori, M. (2008). *Fundamentals of the new Artificial Intelligence* (2ª ed). Springer.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., & Pantic, M. (2016). AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 3–10. https://doi.org/10.1145/2988257.2988258

Velasco, M., Justo, R., & Torres, M. I. (2022). Automatic Identification of Emotional

Information in Spanish TV Debates and Human–Machine Interactions. *Applied Sciences (Switzerland)*, *12*(4). https://doi.org/10.3390/app12041902

Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W. (2022). A systematic review on affective computing: emotion models, databases, and recent advances. In *Information Fusion* (Vols. 83–84). https://doi.org/10.1016/j.inffus.2022.03.009

Yamamoto, M., Takamiya, A., Sawada, K., Yoshimura, M., Kitazawa, M., Liang, K. C., Fujita, T., Mimura, M., & Kishimoto, T. (2020). Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PLoS ONE*, *15*(9 September), 1–10. https://doi.org/10.1371/journal.pone.0238726

Zhou, K., Sisman, B., & Li, H. (2020). *Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data*. https://doi.org/10.21437/odyssey.2020-33

Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion : Theory , databases and ESD. *Speech Communication*, *137*(November 2021), 1–18. https://doi.org/10.1016/j.specom.2021.11.006

Zhu-Zhou, F., Gil-Pita, R., García-Gómez, J., & Rosa-Zurera, M. (2022). Robust Multi-Scenario Speech-Based Emotion Recognition System. *Sensors*, *22*(6), 2343. https://doi.org/10.3390/s22062343

Zloteanu, M., & Krumhuber, E. G. (2021). Expression Authenticity: The Role of Genuine and Deliberate Displays in Emotion Perception. *Frontiers in Psychology*, *11*(January), 1–6. https://doi.org/10.3389/fpsyg.2020.611248

# ANEXOS

## Anexo 1. Aceptación del paciente a realizar el test

**UNIVERSIDAD TÉCNICA DEL NORTE**

**FICA – CITEL**

Un saludo, la Universidad Técnica del Norte (UTN), se encuentra realizando una investigación sobre un: **"**Reconocimiento de emociones a través de la voz, mediante técnicas de aprendizaje profundo**"** por lo que requerimos de su valioso aporte, respondiendo las siguientes preguntas. Nos tomará unos pocos minutos, nuestro compromiso es que toda la información recopilada será trabajada bajo estricto criterio académico y garantizamos absoluta confidencialidad.

Gracias por su tiempo y colaboración.

Edad: ☐ 18 a 25 ☐ 26 a 35 ☐ 35 a 45 ☐ 46 a 55

Género: ☐Femenino ☐Masculino ☐LGBTI ☐Otro

Instrucciones: Este cuestionario consta de 21 grupos de afirmaciones. Por favor, lea con atención cada uno de ellos cuidadosamente. Luego elija uno de cada grupo, el que mejor describa el modo como se ha sentido las últimas dos semanas, incluyendo el día de hoy.

Marque con un círculo el número correspondiente al enunciado elegido Si varios enunciados de un mismo grupo le parecen igualmente apropiados, marque el número más alto. Verifique que no haya elegido más de uno por grupo, incluyendo el ítem 16 (cambios en los hábitos de Sueño) y el ítem 18 (cambios en el apetito).

**Anexo 2. Test de Beck realizado a pacientes**

## UNIVERSIDAD TÉCNICA DEL NORTE

## FICA – CITEL

Un saludo, la Universidad Técnica del Norte (UTN), se encuentra realizando una investigación sobre un: **"**Reconocimiento de emociones a través de la voz, mediante técnicas de aprendizaje profundo**"** por lo que requerimos de su valioso aporte, respondiendo las siguientes preguntas. Nos tomará unos pocos minutos, nuestro compromiso es que toda la información recopilada será trabajada bajo estricto criterio académico y garantizamos absoluta confidencialidad.

Gracias por su tiempo y colaboración.

Edad: ☐ 18 a 25 ☐ 26 a 35 ☐ 35 a 45 ☐ 46 a 55

Género: ☐Femenino ☐Masculino ☐LGBTI ☐Otro

Instrucciones: Este cuestionario consta de 21 grupos de afirmaciones. Por favor, lea con atención cada uno de ellos cuidadosamente. Luego elija uno de cada grupo, el que mejor describa el modo como se ha sentido las últimas dos semanas, incluyendo el día de hoy.

Marque con un círculo el número correspondiente al enunciado elegido Si varios enunciados de un mismo grupo le parecen igualmente apropiados, marque el número más alto. Verifique que no haya elegido más de uno por grupo, incluyendo el ítem 16 (cambios en los hábitos de Sueño) y el ítem 18 (cambios en el apetito).

**1. Tristeza**

0. No me siento triste.

1. Me siento triste gran parte del tiempo

2. Me siento triste todo el tiempo.

3. Me siento tan triste o soy tan infeliz que no puedo soportarlo.

**2. Pesimismo**

0. No estoy desalentado respecto del mi futuro.

1. Me siento más desalentado respecto de mi futuro que lo que solía estarlo.

2. No espero que las cosas funcionen para mí.

3. Siento que no hay esperanza para mi futuro y que sólo puede empeorar.

**3. Fracaso**

0. No me siento como un fracasado.

1. He fracasado más de lo que hubiera debido.

2. Cuando miro hacia atrás, veo muchos fracasos.

3. Siento que como persona soy un fracaso total.

**4. Pérdida de Placer**

0. Obtengo tanto placer como siempre por las cosas de las que disfruto.

1. No disfruto tanto de las cosas como solía hacerlo.

2. Obtengo muy poco placer de las cosas que solía disfrutar.

3. No puedo obtener ningún placer de las cosas de las que solía disfrutar.

**5. Sentimientos de Culpa**

0. No me siento particularmente culpable.

1. Me siento culpable respecto de varias cosas que he hecho o que debería haber hecho.

2. Me siento bastante culpable la mayor parte del tiempo.

3. Me siento culpable todo el tiempo.

**6. Sentimientos de Castigo**

0. No siento que este siendo castigado

1. Siento que tal vez pueda ser castigado.

2. Espero ser castigado.

3. Siento que estoy siendo castigado.

**7. Disconformidad con uno mismo.**

0. Siento acerca de mi lo mismo que siempre.

1. He perdido la confianza en mí mismo.

2. Estoy decepcionado conmigo mismo.

3. No me gusto a mí mismo.

**8. Autocrítica**

0. No me critico ni me culpo más de lo habitual

1. Estoy más crítico conmigo mismo de lo que solía estarlo

2. Me critico a mí mismo por todos mis errores

3. Me culpo a mí mismo por todo lo malo que sucede.

**9. Pensamientos o Deseos Suicidas**

0. No tengo ningún pensamiento de matarme.

1. He tenido pensamientos de matarme, pero no lo haría

2. Querría matarme

3. Me mataría si tuviera la oportunidad de hacerlo.

**10. Llanto**

0. No lloro más de lo que solía hacerlo.

1. Lloro más de lo que solía hacerlo

2. Lloro por cualquier pequeñez.

3. Siento ganas de llorar, pero no puedo.

**11. Agitación**

0. No estoy más inquieto o tenso que lo habitual.

1. Me siento más inquieto o tenso que lo habitual.

2. Estoy tan inquieto o agitado que me es difícil quedarme quieto

3. Estoy tan inquieto o agitado que tengo que estar siempre en movimiento o haciendo algo.

**12. Pérdida de Interés**

0. No he perdido el interés en otras actividades o personas.

1. Estoy menos interesado que antes en otras personas o cosas.

2. He perdido casi todo el interés en otras personas o cosas.

3. .Me es difícil interesarme por algo.

**13. Indecisión**

0. Tomo mis propias decisiones tan bien como siempre.

1. Me resulta más difícil que de costumbre tomar decisiones

2. Encuentro mucha más dificultad que antes para tomar decisiones.

3. Tengo problemas para tomar cualquier decisión.

**14. Desvalorización**

0. No siento que yo no sea valioso

1. No me considero a mí mismo tan valioso y útil como solía considerarme

2. Me siento menos valioso cuando me comparo con otros.

3. Siento que no valgo nada.

**15. Pérdida de Energía**

0. Tengo tanta energía como siempre.

1. Tengo menos energía que la que solía tener.

2. No tengo suficiente energía para hacer demasiado

3. No tengo energía suficiente para hacer nada.

**16. Cambios en los Hábitos de Sueño**

0. No he experimentado ningún cambio en mis hábitos de sueño.

1ª. Duermo un poco más que lo habitual.

1b. Duermo un poco menos que lo habitual.

2a Duermo mucho más que lo habitual.

2b. Duermo mucho menos que lo habitual

3ª. Duermo la mayor parte del día

3b. Me despierto 1-2 horas más temprano y no puedo volver a dormirme

**17. Irritabilidad**

0. No estoy tan irritable que lo habitual.

1. Estoy más irritable que lo habitual.

2. Estoy mucho más irritable que lo habitual.

3. Estoy irritable todo el tiempo.

**18. Cambios en el Apetito**

0. No he experimentado ningún cambio en mi apetito.

1ª. Mi apetito es un poco menor que lo habitual.

1b. Mi apetito es un poco mayor que lo habitual.

2a. Mi apetito es mucho menor que antes.

2b. Mi apetito es mucho mayor que lo habitual

3ª. No tengo apetito en absoluto.

3b. Quiero comer todo el día.

**19. Dificultad de Concentración**

0. Puedo concentrarme tan bien como siempre.

1. No puedo concentrarme tan bien como habitualmente

2. Me es difícil mantener la mente en algo por mucho tiempo.

3. Encuentro que no puedo concentrarme en nada.

**20. Cansancio o Fatiga**

0. No estoy más cansado o fatigado que lo habitual.

1. Me fatigo o me canso más fácilmente que lo habitual.

2. Estoy demasiado fatigado o cansado para hacer muchas de las cosas que solía hacer.

3. Estoy demasiado fatigado o cansado para hacer la mayoría de las cosas que solía

**21. Pérdida de Interés en el Sexo**

0. No he notado ningún cambio reciente en mi interés por el sexo.

1. Estoy menos interesado en el sexo de lo que solía estarlo.

2. Estoy mucho menos interesado en el sexo.

3. He perdido completamente el interés en el sexo.

**Anexo 3. Texto grabado por los pacientes.**

Nos pasamos toda la vida soñando con deseos incumplidos, recordando cicatrices, construyendo artificial y mentirosamente lo que pudimos haber sido. cada vez somos menos verdaderos, más hipócritas; cada vez tenemos más vergüenza de nuestra verdad