Desarrollo de un algoritmo de reconocimiento de emociones utilizando Deep Learning mediante el análisis de la señal de la voz.

Córdova Frías, Dennis Alexander

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Ingeniería en Electrónica y Telecomunicaciones

Trabajo de titulación, previo a la obtención del título de Ingeniero en Electrónica y Telecomunicaciones

Ing. Bernal Oñate, Carlos Paúl, MSc.

26 de agosto del 2021

Curiginal

Document Information

Analyzed document	TRABAJO DE TITULACION DENNIS CORDOVA.pdf (D111674336)
Submitted	8/25/2021 3:56:00 PM
Submitted by	
Submitter email	mgutierrez@difusion.com.mx
Similarity	7%
Analysis address	mgutierrez1.GDC@analysis.urkund.com

Sources included in the report

W	URL: http://repositorio.espe.edu.ec/bitstream/21000/20587/1/T-ESPE-039672.pdf Fetched: 12/18/2020 7:51:49 PM	2
SA	**Trabajo_Titulacion_Tayupanta_Leonardo_Urkund_1.pdf** Document Trabajo_Titulacion_Tayupanta_Leonardo_Urkund_1.pdf (D59153786)	3
W	URL: http://diposit.ub.edu/dspace/bitstream/2445/43135/1/anatomia-funcional-voz.pdf Fetched: 8/25/2021 3:57:00 PM	1
SA	**Flores_Marley_Tesis.pdf** Document Flores_Marley_Tesis.pdf (D54127601)	1
W	URL: https://zenodo.org/record/1188976#.X6vtrGhKjIW Fetched: 8/25/2021 3:57:00 PM	2
W	URL: https://www.sciencedirect.com/science/article/pii/S0167639319302262#bib0139 Fetched: 8/25/2021 3:57:00 PM	2
W	URL: https://ruidera.uclm.es/xmlui/bitstream/handle/10578/26082/TFG_Jose%20Maria%20Guerrero%20Torrijos.pdf?sequence=1&isAllowed=y Fetched: 7/7/2021 2:48:16 AM	3
W	URL: https://www.business2community.com/workplace-culture/the-different-types-of-emotions-and-how-they-impact-human-behavior-02263872 Fetched: 8/25/2021 3:57:00 PM	2
W	URL: https://cacm.acm.org/magazines/2018/5/227191-speech-emotion-recognition/fulltext Fetched: 8/25/2021 3:57:00 PM	2

Firmado electrónicamente por:
CARLOS PAUL
BERNAL ONATE

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que el trabajo de titulación, **"Desarrollo de un algoritmo de reconocimiento de emociones utilizando Deep Learning mediante el análisis de la señal de la voz."** fue realizado por el señor **Córdova Frías, Dennis Alexander** el cual ha sido revisado y analizado en su totalidad por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 26 de agosto de 2021

Firmado electrónicamente por:
CARLOS PAUL BERNAL ONATE

.....

Ing. Bernal Oñate, Carlos Paúl MSc.

C.C 1709775637

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

RESPONSABILIDAD DE AUTORÍA

Yo, **Córdova Frías, Dennis Alexander**, con cédula de ciudadanía n° 1722127212, declaro que el contenido, ideas y criterios del trabajo de titulación: **"Desarrollo de un algoritmo de reconocimiento de emociones utilizando Deep Learning mediante el análisis de la señal de la voz."** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 26 de agosto del 2021

.....
Córdova Frías, Dennis Alexander
C.C.: 1722127212

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Córdova Frías, Dennis Alexander,** con cédula de ciudadanía n° 1722127212, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"Desarrollo de un algoritmo de reconocimiento de emociones utilizando Deep Learning mediante el análisis de la señal de la voz."** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi/nuestra responsabilidad.

Sangolquí, 26 de agosto del 2021

.....
Córdova Frías, Dennis Alexander
C.C.: 1722127212

Dedicatoria

Dedico este trabajo de titulación a mis padres que con tanto esfuerzo me ayudaron y me guiaron es esta carrera.

Dedico a mis profesores que con su apoyo en cada semestre me guiaron para ser un buen profesional

Dedico a mi novia que gracias a su apoyo en los últimos instantes me ayudaron a seguir luchando por este objetivo.

Dennis Alexander Córdova Frías

Agradecimiento

Primero agradezco a Dios y a mis padres por estar en todo este proceso y siempre apoyarme en los momentos más difíciles que se me presentaron a lo largo de la carrera, también agradezco a todas esas personas amigos profesores y conocidos que estuvieron conmigo en todos los semestres a lo largo de la carrera, ya que gracias a ellos pude forjar la persona en la que me convertí, una persona con valores y principios que siempre va a realizar las cosas de forma correcta.

Dennis Alexander Córdova Frías

Índice de contenido

Índice de tablas

Índice de figuras

Resumen

En los últimos años, la tecnología ha sido una clave esencial para que la sociedad obtenga un mejor estilo de vida, con este avance tecnológico existe el método de aprendizaje profundo. Este trabajo de titulación investiga un algoritmo de reconocimiento de emociones mediante el aprendizaje profundo (Deep Learning), utilizando también un modelo de entrenamiento llamado redes neuronales convolucionales.

Este proyecto realizó un clasificador para el reconocimiento de cuatro emociones fundamentales tales como la felicidad, el enojo, el miedo y la tristeza. Para la detección de estas emociones se utilizó una base de datos que contienen 640 audios que se reparte equitativamente entre hombre y mujer. Para el entrenamiento de la red neuronal se utilizó 2 modelos de capas convolucionales uno mediante el uso de 10 capas y otro con el uso de 5 capas, también se utiliza un optimizador de Adam de 32 y de 128, para el desarrollo del proyecto se utilizó el software Matlab®.

Los resultados obtenidos se analizaron mediante cuatro parámetros fundamentales, los cuales son exactitud, precisión, sensibilidad y especificidad, se utilizó el cálculo del Ber para verificar el mejor resultado entre los 5 experimentos que se realizó en cada género.

PALABRAS CLAVE:

- **DEEP LEARNING**
- **RECONOCIMIENTO DE EMOCIONES**
- **CAPAS CONVOLUCIONALES**

Abstract

In recent years, technology has been an essential key for society to obtain a better lifestyle, with this technological advance there is the deep learning method. This degree work investigates an algorithm for the recognition of emotions through deep learning, also using a training model called convolutional neural networks.

This project carried out a classifier for the recognition of four fundamental emotions such as happiness, anger, fear and sadness. For the detection of these emotions, a database containing 640 audios was used that is distributed equally between men and women. For the training of the neural network, 2 models of convolutional layers were used, one using 10 layers and the other with the use of 5 layers, an Adam optimizer of 32 and 128 was also used, for the development of the project it was used Matlab® software.

The results obtained were analyzed using four fundamental parameters, which are accuracy, precision, sensitivity and specificity, the Ber calculation was used to verify the best result among the 5 experiments that were carried out in each gender.

KEYWORDS:

- **DEEP LEARNING**
- **RECOGNITION OF EMOTION**
- **CONVOLUTIONAL LAYERS**

Capítulo I

Introducción del proyecto de Investigación

Antecedentes y justificación del Proyecto

Las emociones que experimentan las diferentes personas en el transcurso de su vida cotidiana es un papel fundamental en los fenómenos sociales, en la actualidad existen estudios que permiten la incorporación del análisis de las emociones a los diferentes objetos de estudio. Las emociones tienen un papel importante en la disciplina de la sociología, al mostrarnos la naturaleza social de las emociones y la infinidad de estados emocionales en una persona (Bericat, 2012).

En las diferentes emociones existe diferentes tipos de complejidades, problemáticas y paradojas, las cuales cuando una persona siente una emoción no se debe considerar como una simple respuesta mecánica o fisiológica. Estas emociones pueden ser causadas por diferentes factores los cuales puede ser como la persona valore conscientemente o inconscientemente algún echo o de que o quien venga ese echo para causar esa emoción (Bericat, 2012).

"La expresión de las emociones en los animales y en el hombre" fue el primer tratado sobre reconocimiento emocional (Chóliz, 1995). En este texto propuesto por Charles Darwin a finales del siglo XIX, afirmaba que existía pruebas sobre la existencia de un mecanismo que ayuda con el reconocimiento de emociones, a partir de la escritura de este texto se observó distintos paradigmas del reconocimiento de emociones tales como la psicología evolucionista, esta psicología es fiel a la teoría de Darwin y trata sobre la adaptación al medio, el otro paradigma es la psicología evolutiva la cual permite destacar el papel de la experiencia y por último la neurociencia este permite un posicionamiento mixto tanto la adaptación al medio como la experiencia, la apertura a

estos paradigmas permitió ingresar al campo del reconocimiento de emociones (Celdrán Baños & Ferrándiz García, 2012) .

Gran parte de las investigaciones que se han realizado acerca del reconocimiento de la emoción del habla ha permitido buscar varias características que indica las diferentes emociones. El enfoque principal que varios investigadores han propuesto es extraer una gran cantidad de características estadísticas tales como, descriptores comunes de bajo nivel (LLD) y funciones estadísticas de alto nivel (HSF) (Mirsamadi et al., 2017).

En 2006, Ververidis y Kotropoulos se enfocaron en la extracción de características del habla, en este mismo tiempo investigaron las características de sonido y los calificaron en una encuesta sobre el reconocimiento de la emoción del habla (Ververidis & Kotropoulos, 2006). Para implementar un sistema de reconocimiento de emociones exitoso, se necesita definir y modelar estas emociones con el fin de que cada emoción tenga su propia característica y permitir su rápido reconocimiento. Para poder tener el reconocimiento de emociones es importante decir que existen estados psicológicos complicados que se debe a diversos componentes. Por lo que decir que existe dos modelos de reconocimiento de emociones del habla es válido, existe dos modelos de reconocimiento de emociones, el modelo emocional discreto y el modelo emocional dimensional. El modelo emocional discreto se basa en las 6 emociones básicas tales como es la tristeza, felicidad, miedo, enojo, asco y sorpresa. El modelo emocional dimensional se basa en utilizar una mínima cantidad de dimensiones latentes para la caracterización de emociones como excitación, control y poder (Berkehan Akça & Oğuz, 2020).

En los años recientes se ha implementado métodos que permiten que los modelos computacionales que están compuestos de múltiples capas de procesamiento tengan un aprendizaje que les permita representar datos con múltiples niveles de abstracción tales como lo es Deep Learning. Estos métodos han ayudado para que obtengan una mejora drástica al implementarlo en el estado del arte para el reconocimiento de la voz, reconocimiento y detección de objetos y varios objetos de estudio que pueda ser implemento por el aprendizaje profundo (LeCun, Bengio, & G. Hinton, 2015).

Este trabajo de investigación permite reconocer las emociones principales de una persona mediante la voz, la cual busca ayudar a una organización que requiera obtener el estado de ánimo de una persona con solo el tono de su voz, el presente trabajo puede ser solicitado por psicólogos, policías, centro de emergencias (call center) y toda aquella organización que permita obtener un beneficio para la sociedad. Con este método de reconocimiento de emociones por voz mejora la atención y eficiencia ante una emergencia que se pude observar y reconocer en un espacio, esto permitirá mejorar la calidad de vida de las personas con trastornos o personas que se encuentran bajo una emergencia.

Objetivos de la Investigación

Objetivo General

- Implementación de un algoritmo de reconocimiento de emociones mediante la utilización del aprendizaje profundo Deep Learning.

Objetivos Específicos

- Elaborar un algoritmo de reconocimiento eficiente que permita tener un acercamiento lo más preciso posible a la emoción expresada por el usuario.
- Investigar los diferentes algoritmos para la clasificación de emociones.

- Entrenar al sistema mediante el aprendizaje profundo Deep Learning a través de la clasificación supervisada.
- Utilizar una base de datos previamente investigada, para la recopilación de datos.
- Analizar evaluar y corregir la precisión del algoritmo de reconocimiento basado en la base de datos aplicada.

Capítulo II

Marco teórico

La voz

La voz que produce nuestro cuerpo se debe a la acción que coordina el cuerpo humano es decir los órganos funcionales que permiten la acción del habla. La producción de la voz humana es una ciencia que implica tanto física como fisiológicamente (Chen, 2016). La voz humana es única en el reino animal por lo que al reflejar la flexibilidad de la voz se puede observar los pensamientos, emociones como alegría, miedo, tristeza etc. Y también se puede observar los pensamientos de las personas a través del habla. La voz es un registro único de cada persona por lo que es una identificación de cada humano.

Estructura general de la voz humana

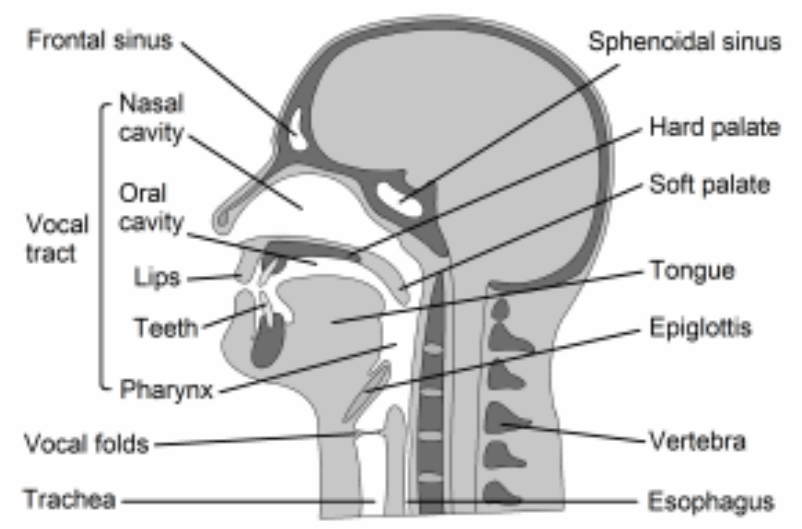
En la Figura 1 se puede observar una sección transversal de los órganos que permiten producir la voz, esta divide el cuerpo en mitades, izquierda y derecha. La fuente de energía que permite la voz es la energía cinética del flujo del aire que produce los pulmones mediante la tráquea. Para producir los efectos de sonidos sonoros, tales como las vocales como [a, e, i, o, u,] y las consonantes como [b, c, d, etc.], una corriente de aire permite que las cuerdas vocales oscilen, permitiendo que el aire se abra y se cierre con cierta frecuencia (Chen, 2016).

El denominado timbre de la voz o color de la voz, está controlada por la faringe, la cavidad oral y la cavidad nasal. La cavidad oral es controlada por diversas partes de los órganos humanos tales como la boca, la lengua, los dientes y los labios, mientras que la

cavidad nasal no tiene una manera de ser controlada, pero es un elemento importante del ser humano para generar la voz (Chen, 2016).

Figura 1

Vista transversal de los órganos de producción de la voz



Nota. La figura muestra la Vista transversal de los órganos de producción de la voz. Tomado de *Elements of Human Voice,* por *(Chen, 2016)*, World Scientific Publishing.

También existe en el cráneo muchas cavidades que se encuentran llenas de aire las cuales funcionan como dispositivos que entran en resonancia por la voz humana. Las cavidades más grandes que existe en el cráneo son el seno frontal y el seno esfenoidal, estas cavidades hacen una conexión directa con la cavidad nasal (Chen, 2016).

Cuerdas vocales

Las cuerdas vocales desempeñan un papel importante en el ser humano que permite producir la voz. Las cuerdas vocales forman un canal de controla el flujo del aire que envían los pulmones, este aire viaja a través de la tráquea y llega al tracto vocal (Chen, 2016).

Las cuerdas vocales son parte del aparato fonador este es el sistema principal de la vibración el cual permite realizar el sonido de la voz, las cuerdas vocales son membranas replegadas de color blanco, existe cuatro cuerdas vocales, dos superiores y dos inferiores, las responsables de generar la voz son las cuerdas vocales inferiores, mientras más separadas se encuentren las cuerdas vocales el tono de la voz será más grave si es que ocurre lo contario y se encuentran más cercanas estas producirán una voz más aguda.

Laringe

La laringe es un órgano tubular que permite la protección de las vías respiratorias y produce sonidos mediante el sistema respiratorio. La laringe es resistente y flexible y permite conectar la faringe con la tráquea en el cuello, la laringe realiza un papel fundamental en el tracto respiratorio, este órgano ayuda a que los alimentos o bebidas no bloqueen el tracto respiratorio y que el aire realice su paso normal, este órgano es responsable de la caja de la voz, debido a las cuerdas vocales que permiten los sonidos del habla (Barclay, 2019).

Glotis

La glotis es una parte de la laringe la cual es la responsable de producir la voz, esta incluye los pliegues vocales y el espacio entre ellas, la cual se llama hendidura glótica. La voz se produce cuando la exhalación del aire pasa a por la glotis (hendidura glótica), las cuerdas vocales se ven obligadas a separarse y se ponen en vibración debido a la presión que produce el aire (Torres, 2007).

Postura del pliegue vocal

Para que la voz se pueda comunicar se requiere que exista un control fino y el ajuste del tono, volumen y calidad de la voz. El ajuste de estos componentes se realiza mediante la activación del musculo laríngeo, este musculo permite que las cuerdas vocales se endurezcan, deformen o reposicionen (Zhang, 2016).

Mediante el movimiento de los cartílagos aritenoides, se logra una postura de aducción y abducción de las cuerdas vocales, el movimiento de los cartílagos aritenoides se debe a la articulación cricoaritenoides que permite que se deslicen y giren alrededor del eje largo del cartílago cricoides. Los músculos cricoaritenoides son los responsables de abrir la glotis, lo cual realiza un papel fundamental para producir la voz en tonos muy altos (Zhang, 2016).

El sonido de la voz se produce debido a que los fenómenos aerodinámicos permiten que las cuerdas vocales vibren de tal manera que realiza secuencias de ciclos vibratorios, estos ciclos vibratorios se dividen de tal manera:

- Tono bajo: 110 ciclos por segundo (Hombre)
- Tono medio: 180 a 220 ciclos por segundo (Mujer)
- Tono alto: 300 ciclos por segundo (Niños)

Ciclo vibratorio

El ciclo vibratorio se compone de una fase abierta más una fase cerrada, esto quiere decir que las cuerdas vocales incluyen una secuencia bien ordenada que se abre y se cierra tanto en la parte superior como en la parte inferior de las cuerdas vocales, esto permite el paso de pequeñas bocanadas de aire. La columna permite realizar un efecto de Bernoulli para el paso del aire y con esto controla la fase de cierre (The voice fundation, 2017).

Tracto vocal

Los resonadores y articuladores (nariz, faringe y boca), amplifican y modifican el sonido que realiza la voz, esto permite obtener una cualidad única de la voz en cada persona. La forma en que se produce la voz es similar a la que realiza un trombón. El trombón produce el sonido cuando se envía el aire por la boquilla del mismo lo cual permite que el aire vibre y pase por la boca. Mientras se cambie el deslizamiento del trombón el sonido cambiará al igual que las cuerdas vocales en las personas (The voice fundation, 2017).

Proceso de la voz

1.- La columna de presión de aire se mueve hacia las cuerdas vocales.

2.- Mediante el diafragma, el abdomen, el pecho y la caja toráxica el aire pasa desde los pulmones hasta las cuerdas vocales.

3.- Secuencia de ciclo vibratorio y la vibración del pliegue vocal, los pliegues vocales se mueven a la línea media de los músculos, nervios y cartílagos.

Mientras que el ciclo vibratorio se produce de la siguiente manera:

- La columna de presión de aire abre la parte inferior de las cuerdas vocales.
- La columna de aire se sigue moviendo mientras se desplaza hacia la parte superior de las cuerdas vocales.
- La presión que produce la columna de aire realiza un efecto de Bernoulli y permite que la parte inferior se cierre y después se cierre la parte superior

- Al cerrar las cuerdas vocales permite que se corte el paso del aire en la columna y libere un pulso de aire.
- Y se repite varias veces este proceso.

Los pulsos rápidos creados por la repetición de este ciclo vibratorio producen un sonido sonoro, lo que en realidad solo es un zumbido, pero luego es modificado y amplificado por los resonadores del tracto vocal, el sonido que realiza después de este proceso se lo conoce como voz (The voice fundation, 2017).

Las emociones

Definición de las emociones

Las emociones son un proceso del organismo humano que se activa mediante la detección de alguna amenaza, peligro o desequilibrio, esto permite poner en acción los recursos de cada persona para poner al frente alguna situación. Cada individuo puede experimentar una emoción de forma diferente debido a las experiencias que viven o la forma en la que tienen un dialogo con alguna otra persona. A veces las personas pueden ser gobernadas por las emociones y estas hacen que se tomen acciones positivas o negativas hacia uno mismo o hacia otra persona, es importante tener un claro control sobre estas emociones (Morzaria, 2019).

Clasificación de las emociones

Hay cuatro emociones básicas que experimenta una persona, es independiente de su cultura o raza. Estas emociones son felicidad, tristeza, ira y miedo.

Felicidad

La felicidad es una de las cuatro emociones básicas, la felicidad es un estado emocional encantador y su característica principal se observa en el sentimiento de alegría. La felicidad es fácil de reconocer a través de expresiones corporales o expresiones faciales como la risa o la voz dulce (Morzaria, 2019).

Tristeza

La tristeza es un estado emocional pasajero, normalmente se debe a sentimientos de pena, desilusión desesperanza y desinterés. Esta emoción la experimenta todas las personas, en ocasiones esta emoción dura largos periodos, hasta se puede convertir en depresión. La tristeza se puede expresar mediante la tranquilidad de una persona o el estado de ánimo disminuido, el letargo o el llanto (Morzaria, 2019).

Ira

Esta emoción es muy importante ya que se caracteriza por tener sentimientos de frustración, hostilidad, agitación hacia uno mismo o hacia otras personas. Esta emoción suele aparecer cuando alguna cosa no sale como lo queremos o cuando la persona se siente amenazado por otra persona o por alguna cosa (Morzaria, 2019).

Miedo

La emoción del miedo es una emoción muy importante en el ser humano ya que juega un papel esencial en la supervivencia del mismo. Con la emoción del miedo una persona puede huir o pelear cuando tiene la sensación de miedo. Cuando se experimenta miedo el ritmo cardiaco aumenta, los músculos se tensan y su mente se mantiene alerta. (Morzaria, 2019).

El reconocimiento de las emociones

La comunicación con inteligencia artificial cada vez se a convertido en algo real, tal es el caso como Alexa, Cortana, Siri y otros sistemas de diálogo que ha logrado llegar al alcance de las personas, esto permite tener un diálogo con esta inteligencia artificial y que poco a poco mejore este sistema hasta convertirse en un compañero de conversación inteligente (Schuller, 2018).

El reconocimiento automático de emociones humanas y las expresiones se las conoce como *Speech Emotion Recognition* (SER). Estudios realizados con anterioridad en la Psicología han permitido realizar una investigación en el papel de la acústica de la emoción humana. Blaton escribió “Todas las personas reconocen el efecto de las emociones en la voz, los más primitivos pueden reconocer los tonos del amor, el miedo y este conocimiento lo comparten los animales. El perro, el caballo y muchos otros animales pueden comprender el significado de la voz humana. El lenguaje de los tonos es el más antiguo y universal de todos nuestros medios de comunicación” (Schuller, 2018).

Características

La extracción de características juega un papel importante para el reconocimiento de emociones, la representación general de las características permite obtener un enfoque acertado del tipo de emoción a reconocer (Anton Batliner, 2011).

Características acústicas

Al hablar de reconocimiento de emociones se debe hablar del modelamiento de unidades de análisis más largas como enunciados o giros o los movimientos de diálogos. Para poder realizar este objetivo las características que se pueden obtener deben ser en

forma de cuadro y se puede combinar mediante medidas apropiadas, como por ejemplo el promedio o también se recurre a una clasificación dinámica.

Los tipos de características acústicas son:

- *Duración:* Se lo puede diferencia por la presentación de aspectos temporales del modelo, la unidad básica de la duración es el milisegundo (ms). Se puede aplicar en distintos tipos de normalización. Los atributos de duración se observan con respecto a la naturaleza de extracción, depende de la representación de aspectos temporales de otros contornos de base acústica y de la representación exclusiva del parámetro duración de unidades fonológicas tales como los fonemas, las sílabas, las palabras, pausas y enunciados (Anton Batliner, 2011).
- *Intensidad:* En la intensidad generalmente se modela el volumen de un sonido, tal y como lo percibe un humano, esta se basa en la amplitud de diferentes intervalos. Aquí se aplica diferentes tipos de normalización. La intensidad modela intervalos o características de puntos. Con el aumento de la intensidad la sensación del oído aumenta logarítmicamente, las unidades son los decibelios. Es importante recalcar que la percepción de sonido depende de la distribución espectral y de la duración (Anton Batliner, 2011).
- *Pitch:* La extracción del tono ha resultado de gran utilidad ya que casi todos los algoritmos de detección de tono (PDA) se obtiene mediante el análisis de cuadros. La señal de la voz se divide en cuadros superpuestos y se obtiene un valor de tono para cada segmento por autocorrelación. Existe otros métodos que permite utilizar la representación cepstral que es básicamente la

explotación de la información armónica implementando la compresión espectral (Anton Batliner, 2011).

- *Espectro:* Se caracteriza por máximos espectrales (formantes) que permite modelar el contenido hablado, los formantes superiores también permite obtener este modelamiento. El espectro y los formantes se representa mediante la posición, la amplitud y el ancho de banda (Anton Batliner, 2011).
- *Calidad de voz:* Esta característica representa el Jitter, el brillo y otros modelos micro prosódicos. Existen otros modelos de calidad de oz como la relación a ruido armónico (NHR) o la relación armónica de ruido (HNR) (Anton Batliner, 2011).
- *Wavelets:* Esta técnica acústica permite analizar una resolución múltiple a corto tiempo, también analiza la energía y las frecuencias de la señal de voz. Esta característica es superior en el modelado de aspectos temporales.

Deep Learning

Deep Learning se puede definir como una técnica avanzada de inteligencia artificial que conlleva una idea de la imitación del cerebro humano mediante hardware y software, utiliza abstracción jerárquica, es decir, representa datos en varios niveles o en capas de aprendizaje automático, las cuales permite extraer características de un nivel de complejidad mayor y procesarlo para su aprendizaje (Arteaga, 2015).

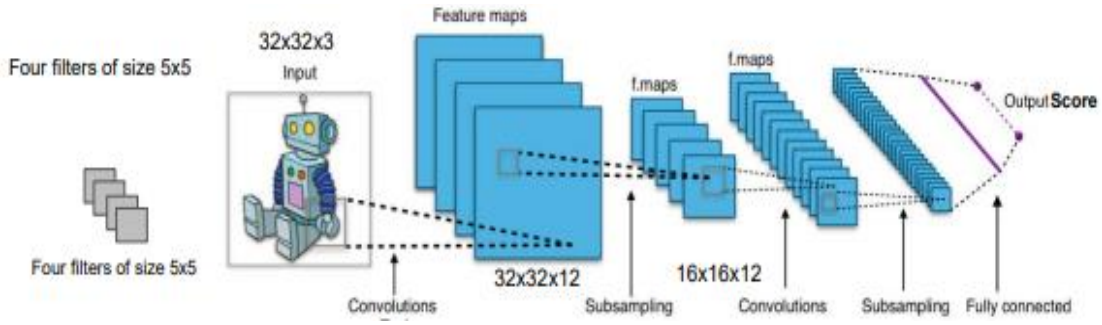
Los primeros indicios de la aparición de Deep Learning son en el año 2006, cuando Geoffrey Hinton añade este término para explicar nuevas arquitecturas de redes neuronales que permite mejorar el aprendizaje de modelos planos. En la actualidad la inteligencia artificial ya no es un mito, si no se ha vuelto una realidad que ha permitido el desarrollo de nuevas tecnologías (Hernández et al., 2018).

Redes Neuronales Convolucionales

La red neuronal convolucional (CNN convolutional neural network) es un tipo de red neuronal artificial que permite el reconocimiento y procesamiento de imágenes en 2D, como se puede observar en la Figura 2, se tiene una entrada de imágenes, esta red realiza un proceso mediante las siguientes capas tales como la capa convolucional, la capa de submuestreo, capa de agrupación, capa totalmente conectadas y capas de normalización. Esta red neuronal es un procesador de imágenes muy fuerte que mediante el aprendizaje profundo (Deep Learning) permite extraer características descriptivas mediante el reconocimiento de imágenes y el procesamiento del lenguaje natural (Rouse, 2018).

Figura 2

Ejemplo de una Red neuronal convolucional



Nota. La figura muestra el ejemplo de una Red neuronal convolucional (CNN). Tomado de *Introduction to Deep Learning*, por Instrument *(2018)*, Training.ti.

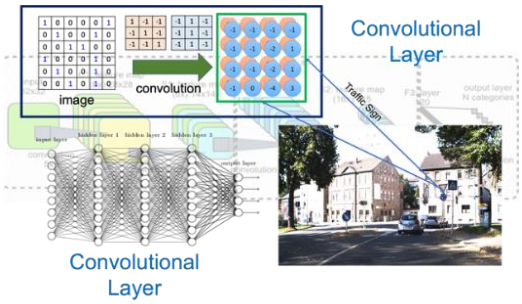
Capa Convolucional

La capa convolucional es una capa que contiene un conjunto de filtros, las cuales tienen características que necesitan ser aprendidas y cuyos bloques son los principales para poder construir una red convolucional neuronal como se puede observar en la Figura 3. Este conjunto de filtros o también llamados *kernels,* tienen la función de realizar una operación de convolución para la primera capa, toma un subconjunto de los datos de entrada, las operaciones que realiza en esta capa son multiplicaciones lineales, el objetivo principal es la extracción de características de alto nivel. (SHARMA, 2020).

Una vez extraído las características de la capa anterior, realiza un mapeo de su apariencia, y los lleva a un mapa de características. La capa realiza una convolución de la entrada, mueve los filtros a lo largo de la entrada vertical y horizontal, calculando el productor escalar y los pesos de la entrada.

Figura 3

Capa Convolucional



Nota. La figura muestra la Capa Convolucional Layer. Tomado de *Role of Convolutional Layer in Convolutional Neural Network*, por Sharma (2020), Vinod Sharma's Blog.

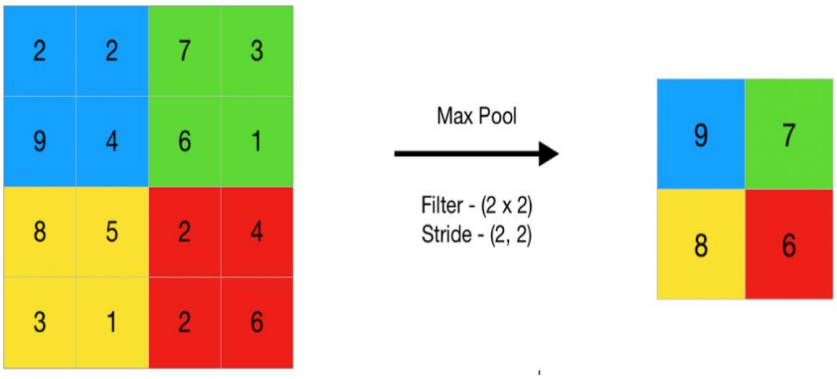
Capa de agrupación

La capa de agrupación permite agrupar las características que se obtiene en una parte de una región que contiene el filtro generado por la capa de convolución, en la Figura 4 se puede observar cómo permite reducir el tamaño de los mapas de entidades, al implementar la capa de agrupación después de la capa convolucional permite tener un orden de capas ya que se pueden repetir en una red neuronal convolucional (Savyakhosla, 2021).

La capa de agrupación máxima permite realizar una operación de agrupación que selecciona un elemento máximo dentro del mapa de características. El resultado de esta agrupación es un nuevo mapa con características destacadas del mapa de características principal.

Figura 4

Max Pooling



Nota. La figura muestra el Max Pooling. Tomado de *Introduction to Pooling Layer*, por Savyakhosla, *(2021)*, GeeksforGeeks.

Capítulo III

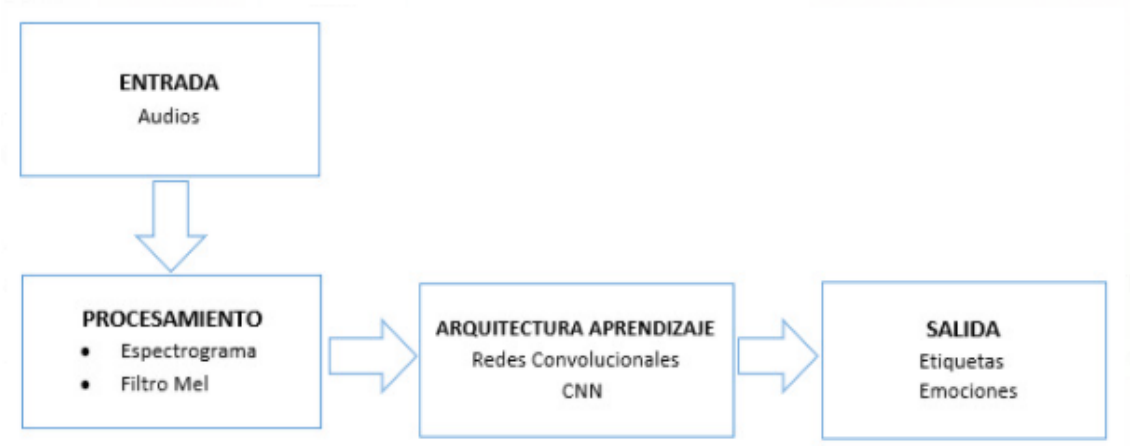
Metodología del proyecto de investigación

Descripción general del proyecto de Investigación

El presente trabajo de investigación permite el reconocimiento de cuatro emociones básicas tales como la felicidad, el enojo, la tristeza y el miedo, mediante el análisis de la señal de voz de hombres y de mujeres utilizando la herramienta de software Matlab®, este software permitirá extraer características utilizando el procesamiento de señales, el método para el entrenamiento del sistema es Deep Learning el cual es una técnica de aprendizaje supervisado.

Figura 5

Diagrama de bloques del Proyecto



En la Figura 5 se puede observar el diagrama de bloques del proyecto, en primer lugar, se tiene un bloque el cual permite el ingreso de los datos mediante audios, los cuales son procesados utilizando los filtros Mel y después se grafica los respectivos espectrogramas, en el siguiente bloque se puede observar el modelo utilizado para el

entrenamiento de la red, mediante redes convolucionales CNN y finalmente se tiene los resultados que es el etiquetado de las emociones.

Entrada de los datos de audio

En la entrada de datos se detalla los audios que se pudo encontrar en la base de datos de “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)” (Livingstone & Russo, 2018) la cual permite la descarga de 80 audios por cada emoción y por cada género como se va a describir en la siguiente sección.

Base de datos (RAVDESS)

Los audios presentan las siguientes características

- Velocidad de bits 390kbps
- Frecuencia 16KHz
- Duración 3 segundos

La base de datos para los primeros 4 experimentos consta de:

- Enojo Hombre y Enojo Mujer: Tiene 80 audios por cada género.
- Feliz Hombre y Feliz Mujer: Tiene 80 audios por cada género.
- Miedo Hombre y Miedo Mujer: Tiene 80 audios por cada género.
- Triste Hombre y Triste Mujer: Tiene 80 audios por cada género.

La base de datos para el quinto experimento consta de:

- Enojo Hombre y Enojo Mujer: Tiene 13 audios por cada género.
- Feliz Hombre y Feliz Mujer: Tiene 13 audios por cada género.
- Miedo Hombre y Miedo Mujer: Tiene 13 audios por cada género.
- Triste Hombre y Triste Mujer: Tiene 13 audios por cada género.

Para el entrenamiento con falsos positivos se tuvo audios de otras emociones tales como sorpresa, disgusto. En los cuales también había 80 audios por cada emoción y por cada género para los primeros 4 experimentos y para el quinto experimento un total de 13 audios por cada emoción.

El entrenamiento de la red neuronal se dividió en dos segmentos, un segmento de testing y otro de validation, el cual se dividió en un 80% para *testing* y un 20% para *validation*, también se realizó una clasificación para el entrenamiento mediante el cambio del Optimizador de Adam se cambió entre dos valores uno de 32 y otro 128 para realizar el respectivo entrenamiento con estos valores.

Para el quinto experimento se utiliza un total de 13 audios, el cual es la Base de Datos Original en la tesis de la Srita. Evelyn Flores (MARLEY, 2019), en este entrenamiento se dividió en un 60 % para *testing* y un 40 % para *validation,* de igual manera se realizó la clasificación mediante el cambio de Optimizador de Adam a 128.

El Optimizador de Adam es un proceso de gradiente estocástico para el entrenamiento de modelos de Deep Learning, este algoritmo permite la optimización necesaria para el manejo adecuado de gradientes dispersos en problemas ruidosos (Kingma & Ba, 2014).

Procesamiento del Audio

El proceso que se utilizó para convertir el audio en imágenes es mediante el uso de espectrogramas, el método que se utiliza es el Espectrograma Mel, el Espectro Mel es un espectrograma en la cual las frecuencias se convierten a la Escala Mel, la escala Mel es una propuesta por los científicos Stevens Volkmann y Newmann es una escala perceptiva de tonos que los oyentes considera iguales en distancia entre sí, esta escala

permite representar digitalmente una señal de audio, lo que se realiza primero es dividir la señal de audio en tramas y una función de ventaneo el cual en este caso es una ventana de Hamming de 25ms, lo que permite esta ventana es eliminar los bordes de la señal y resaltar el centro de la trama. (ICHI.PRO, 2020-2021).

Mediante la Transformada rápida de Fourier (STFT) se obtiene la amplitud del espectro y esta información pasa al dominio Mel mediante un banco de filtros. Para mejorar estos datos y obtener una distribución más uniforme se utiliza un desplazamiento llamado epsil de 1e-15. En las Figuras 6, 7, 8 y 9 se puede observar los diferentes espectrogramas de cada emoción que se genera mediante este proceso.

Espectrogramas Hombres

Figura 6

Espectrogramas Hombres

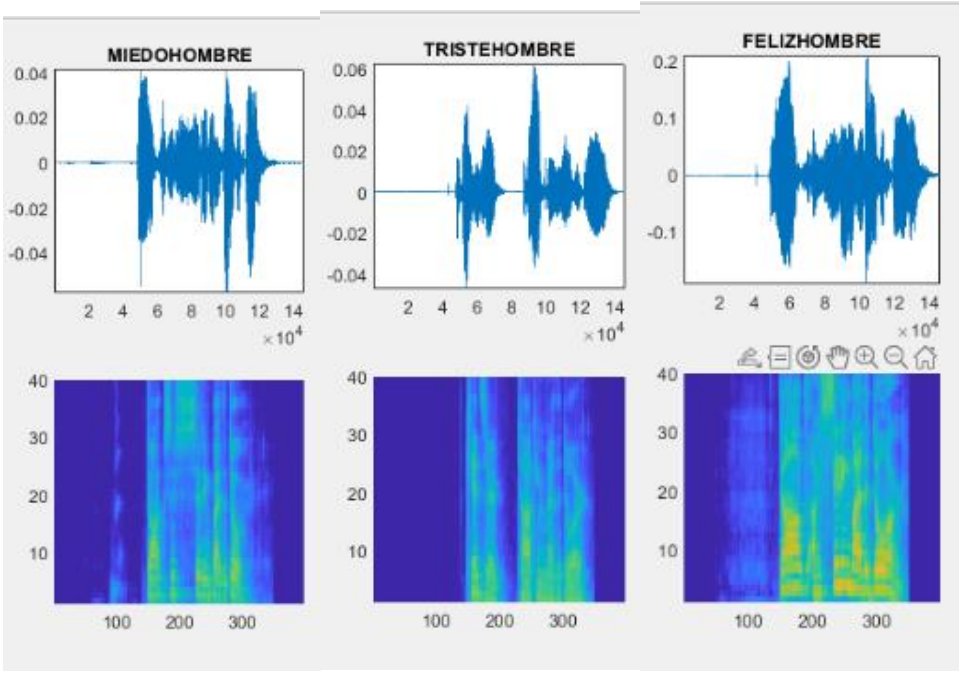
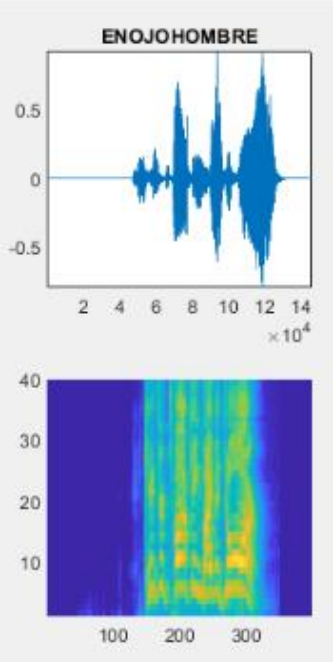


Figura 7

Espectrogramas de las Emociones Hombres



En las Figuras 6 y 7 se puede observar el análisis del espectrograma de un audio por cada emoción, se puede observar también que el espectrograma utiliza la escala Mel, como se explica en la sección 3.3, el espectrograma Mel elimina los bordes de la señal y resalta el centro de la trama para su correcto procesamiento.

Espectrogramas Mujeres

Figura 8

Espectrogramas Mujeres

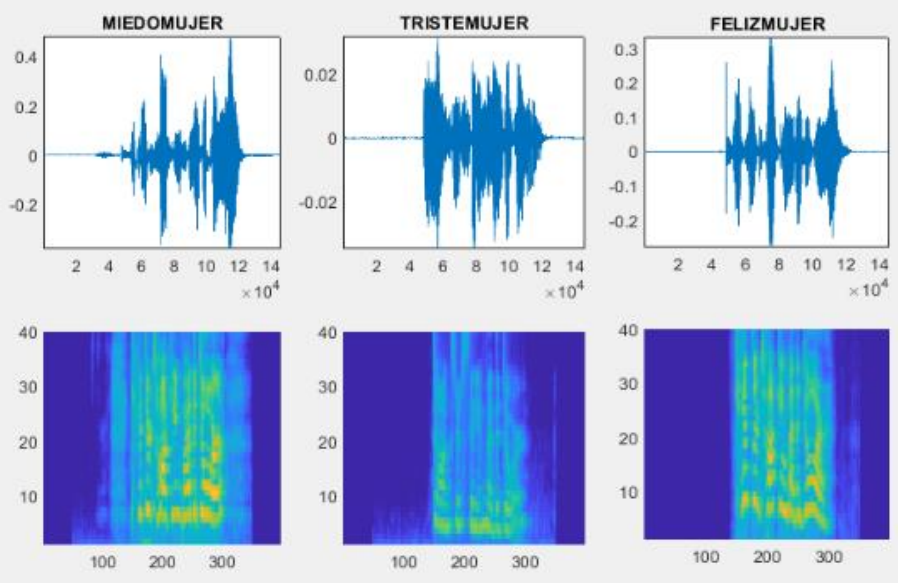
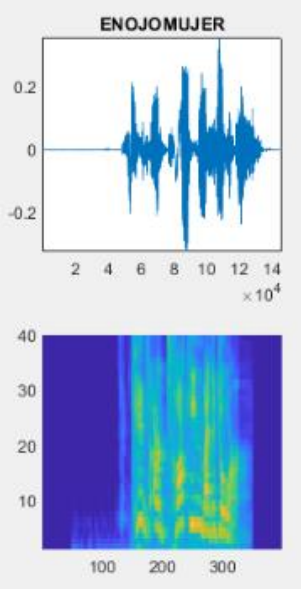


Figura 9

Espectrogramas de las Emociones Mujeres



En las Figuras 8 y 9 se puede observar el análisis del espectrograma de un audio por cada emoción, se puede observar también que el espectrograma utiliza la escala mel, como se explica en la sección 3.3, el espectrograma mel elimina los bordes de la señal y resalta el centro de la trama para su correcto procesamiento.

Arquitectura de la red convolucional

La arquitectura que se propone se divide en 2 partes con diferentes resultados el primer modelo es de 10 capas convolucionales y otra de 5 capas convolucionales, el primer modelo de 10 capas convolucionales utiliza un orden de capas convolucionales CR-CR-CRPB-CR-CR-CRPB-CR-CR-CRPB-CRPFS y el segundo modelo utiliza un orden de capas convolucionales CR-CR-CRPB-CR-CR- CRPFS, se escogió este modelo porque es el modelo más óptimo que se obtuvo con los diferentes experimentos realizados, para el entrenamiento se verificó con varios modelos de redes convolucionales pero no mejoraba el entrenamiento y como resultado presentaba muchos errores tanto en el entrenamiento como en la validación, a continuación, se puede validar el significado y funcionalidad de cada capa.

- C: convolution2dLayer
- B: batchNormalizationLayer
- P: maxPooling2dLayer
- R: reluLayer,
- F: fullyConnectedLayer

Convolution2Layer

La capa convolucional 2-D, permite aplicar filtros convolucionales en la entrada. Esta capa permite la convolución de la entrada moviendo los filtros de la entrada horizontal y

vertical. (MathWorks, convolution2dLayer, 2021). Esta capa convolucional trabaja con 48 filtros de tamaño [5 5], con un momento de relleno para que la salida de la capa tenga la misma salida que la entrada.

BatchNormalizationLayer

La capa de normalización por lotes utiliza un pequeño lote de datos para cada canal de manera independiente. Esto permite acelerar el entrenamiento de la CNN y también permite reducir la sensibilidad a al inicio de la red (MathWorks, batchNormalizationLayer, 2021).

MaxPooling2dLayer

La capa de agrupación máxima permite realizar un muestreo reducido, esta capa permite dividir la entrada en regiones rectangulares y calcula el máximo de cada región (MathWorks, maxPooling2dLayer, 2021). Esta capa tiene un tamaño de 5 filtros y un *stride* o paso de 3x3.

ReluLayer

La capa de unidad línea rectificada permite realizar una operación de umbral el cual en cada elemento de entrada que tiene un valor menor a cero, se establece el valor de cero (MathWorks, reluLayer, 2021).

FullyConnectedLayer

La capa completamente conectada permite multiplicar la entrada por una matriz de peso y luego esta agrega un vector de polarización (MathWorks, fullyConnectedLayer, 2021).

Entrenamiento y Validación de la red

Como se explica en la sección 3.2.1 y 3.3, obteniendo los espectrogramas se realiza el entrenamiento de la red neuronal mediante el optimizador de Adam (Kingma & Ba, 2014), variando entre dos valores entre 32 y 128, también se utiliza un máximo de épocas de 25, se realizó un incremento de épocas, pero en el entrenamiento de la red neuronal los datos de validación se alejaban mucho a las de entrenamiento, esto producía valores más altos de error, después se redujo la tasa de aprendizaje en un factor de 10 luego de 20 épocas, todos estos parámetros permiten que los valores de error en entrenamiento y validación sean mínimos.

A continuación, se puede observar las gráficas obtenidas de los entrenamientos de la red neuronal y los resultados obtenidos mediante las matrices de confusión de los experimentos realizados.

Red Neuronal y Matriz de confusión Hombres

Figura 10

Experimento 1

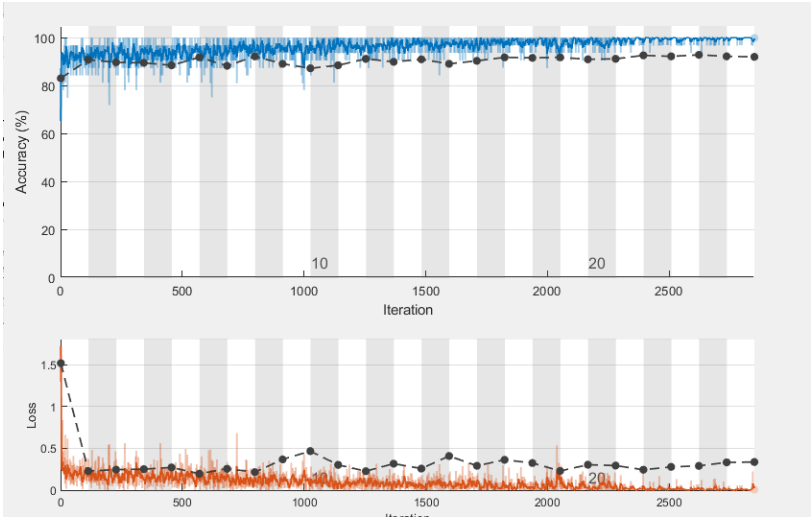


Figura 11

Experimento 1

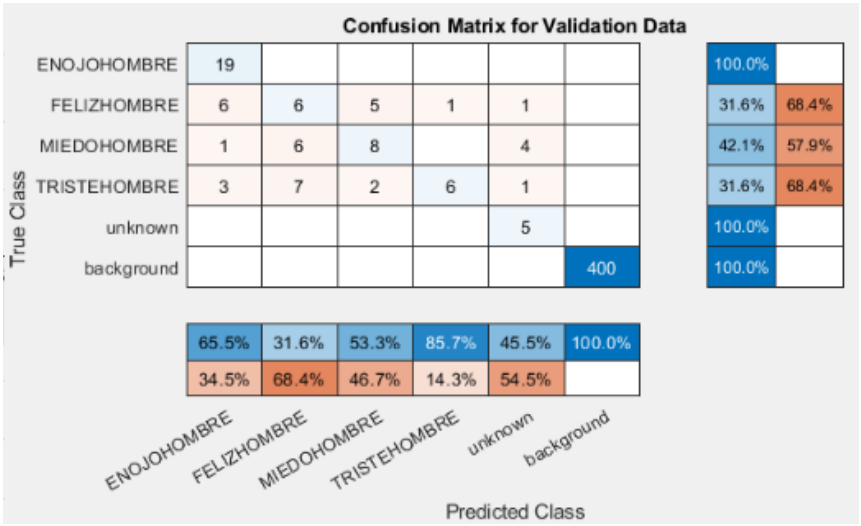


Figura 12

Experimento 2

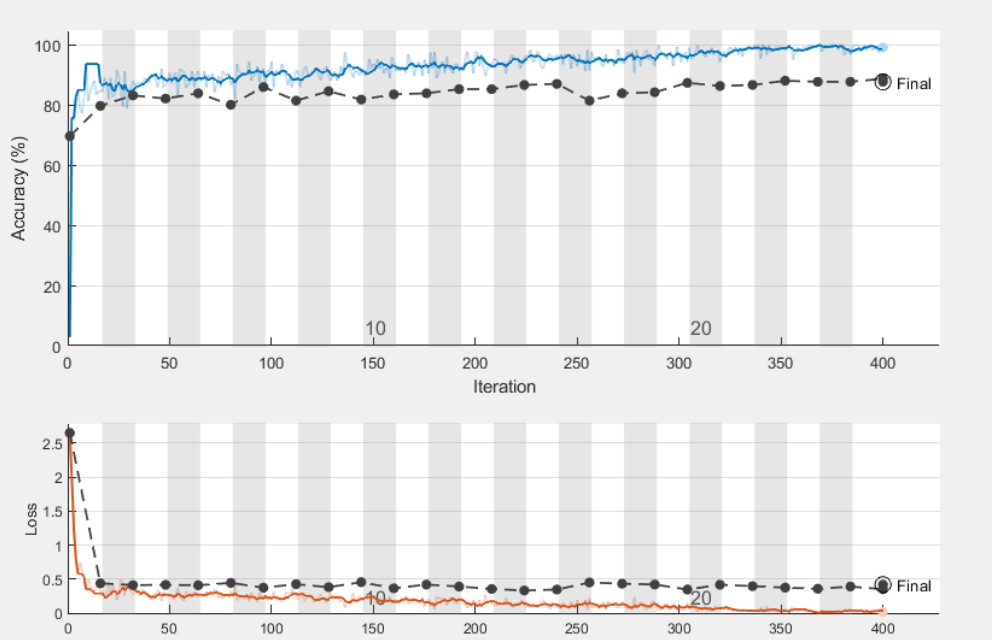


Figura 13

Experimento 2

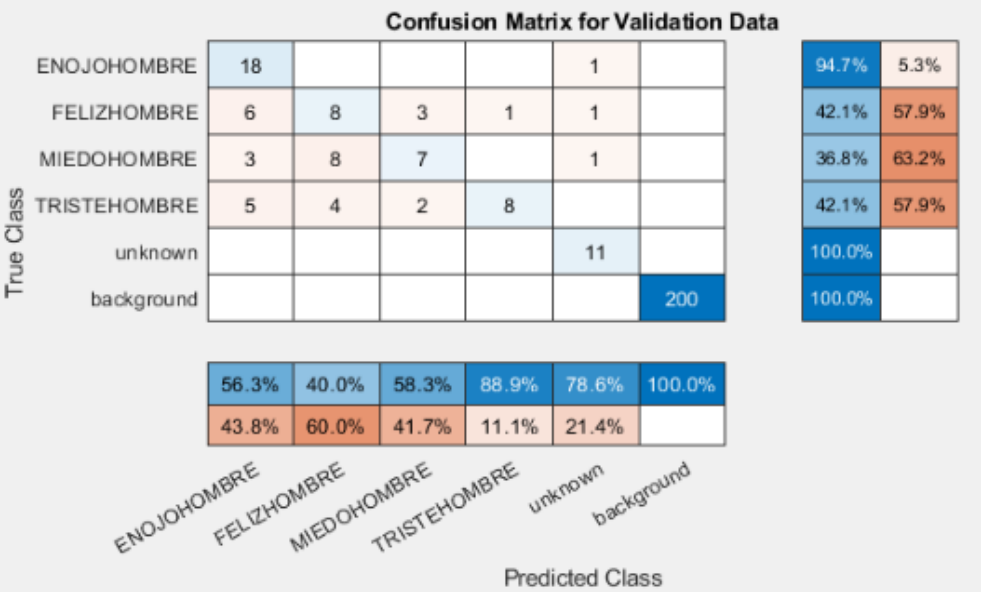


Figura 14

Experimento 3

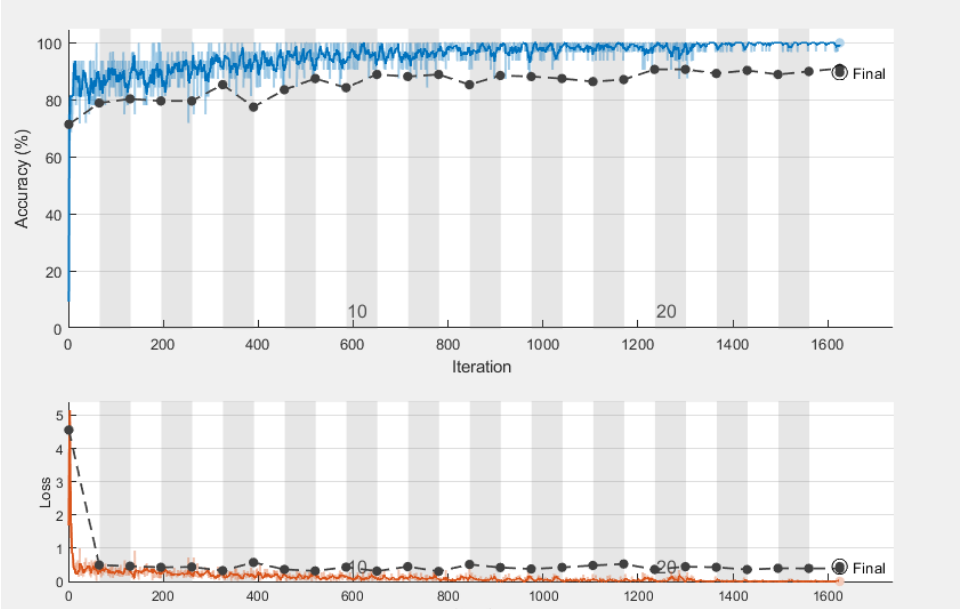


Figura 15

Experimento 3

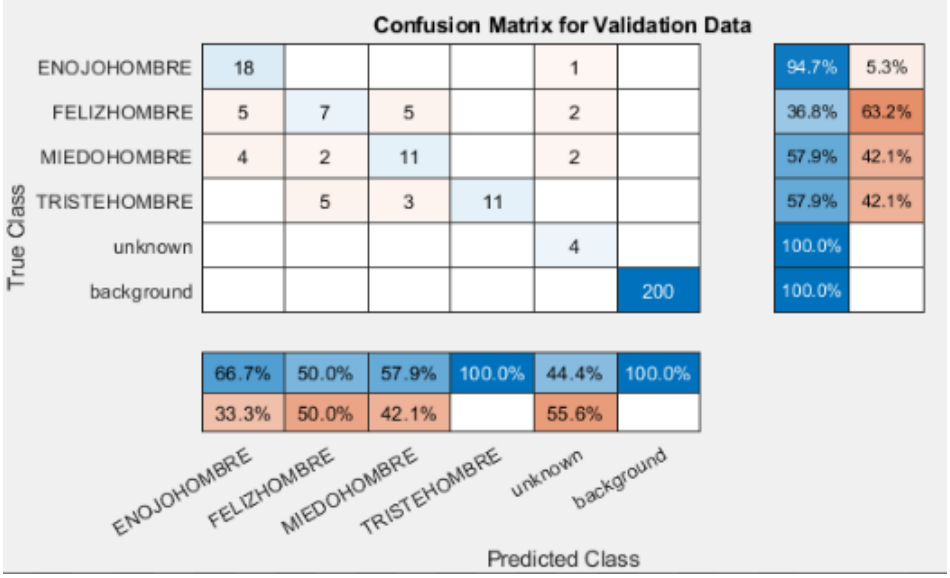


Figura 16

Experimento 4

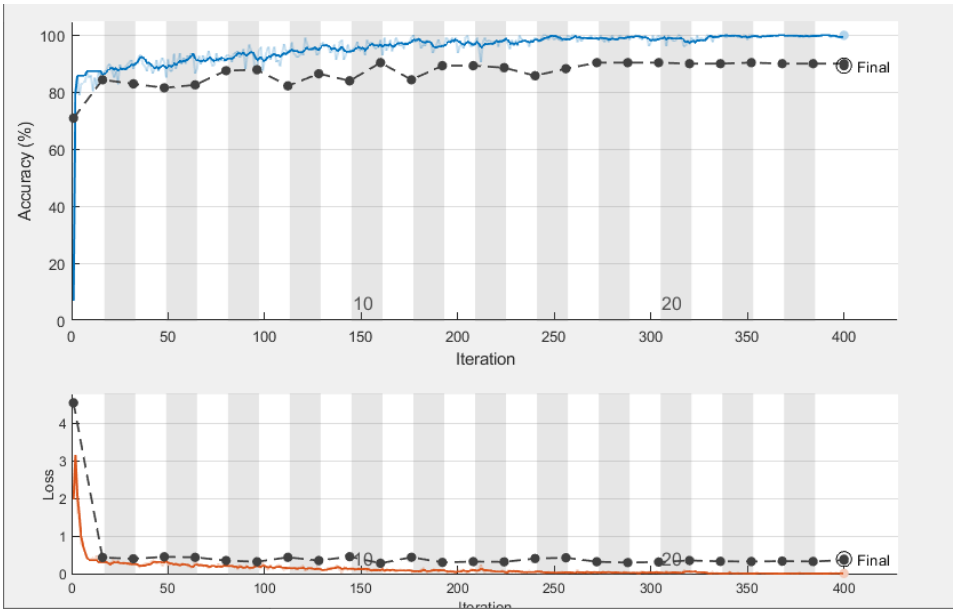


Figura 17

Experimento 4

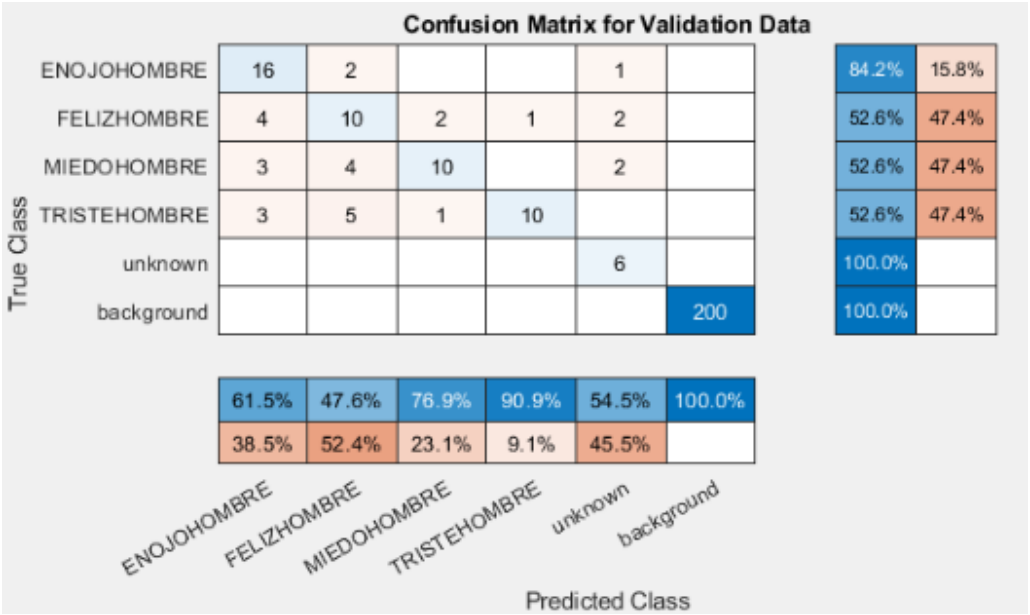


Figura 18

Experimento 5

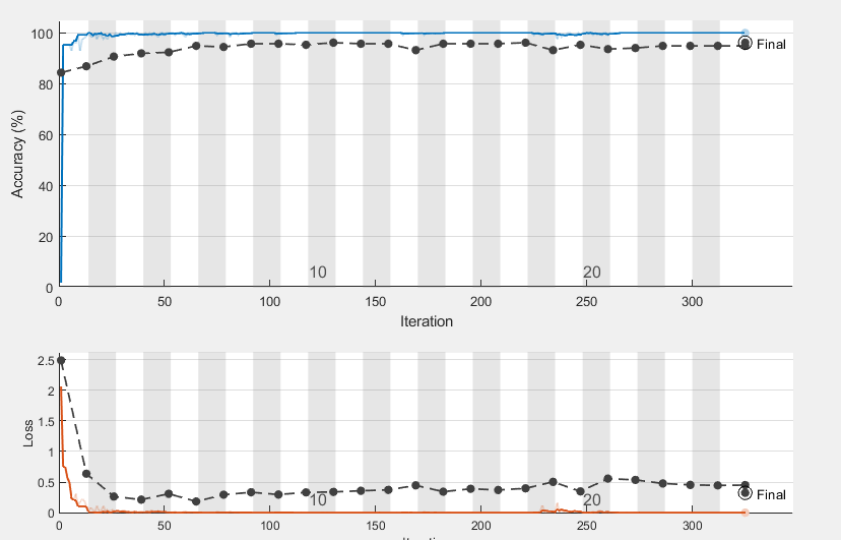
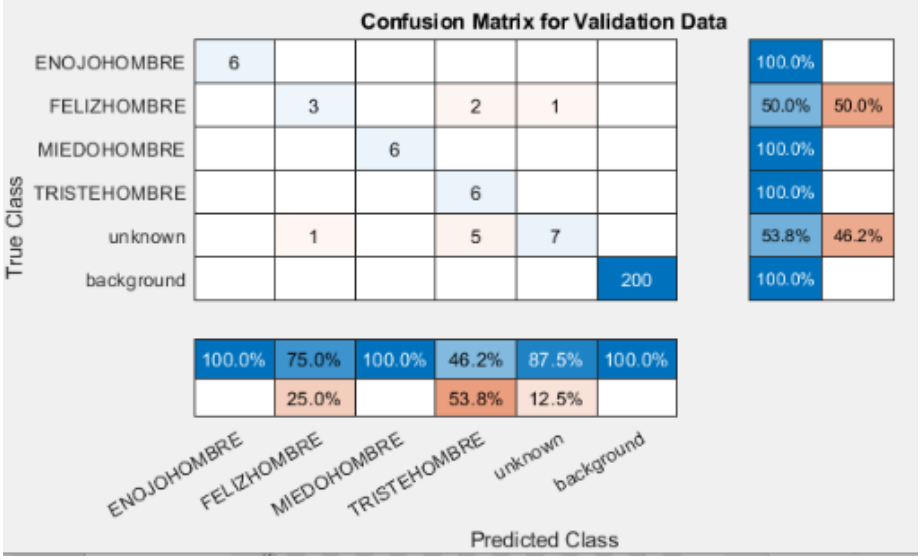


Figura 19

Experimento 5



En las Figuras 10, 12, 14, 16, 18 se puede observar el entrenamiento de la red convolucional en donde la línea azul representa la exactitud que tiene el entrenamiento de la red neuronal, la línea naranja representa las pérdidas que se tiene en la red neuronal y la línea negra presenta la validación tanto en exactitud como en pérdidas.

En las Figuras 11, 13, 15, 17, 19 se puede ver la matriz de confusión en donde se observa el porcentaje de aciertos y errores por cada emoción, con estos datos se puede obtener los resultados de exactitud, precisión, sensibilidad y especificidad.

Red Neuronal y Matriz de confusión Mujeres

Figura 20

Experimento 1

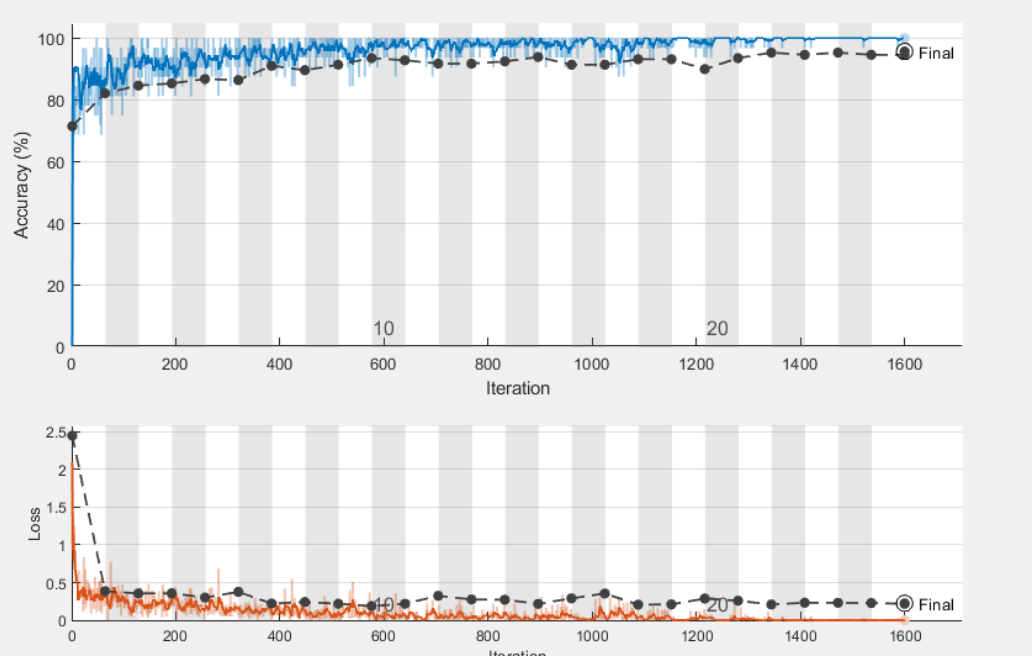


Figura 21

Experimento 1

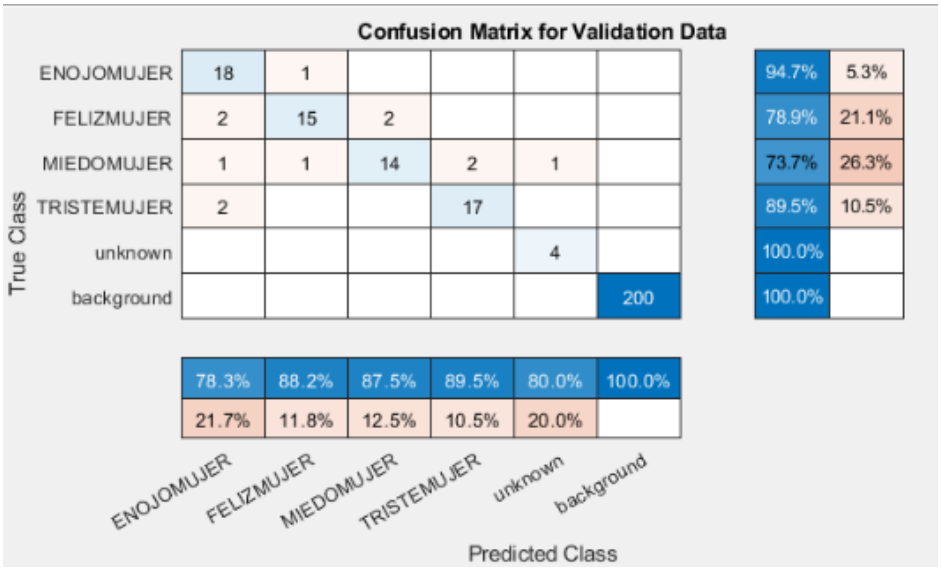


Figura 22

Experimento 2

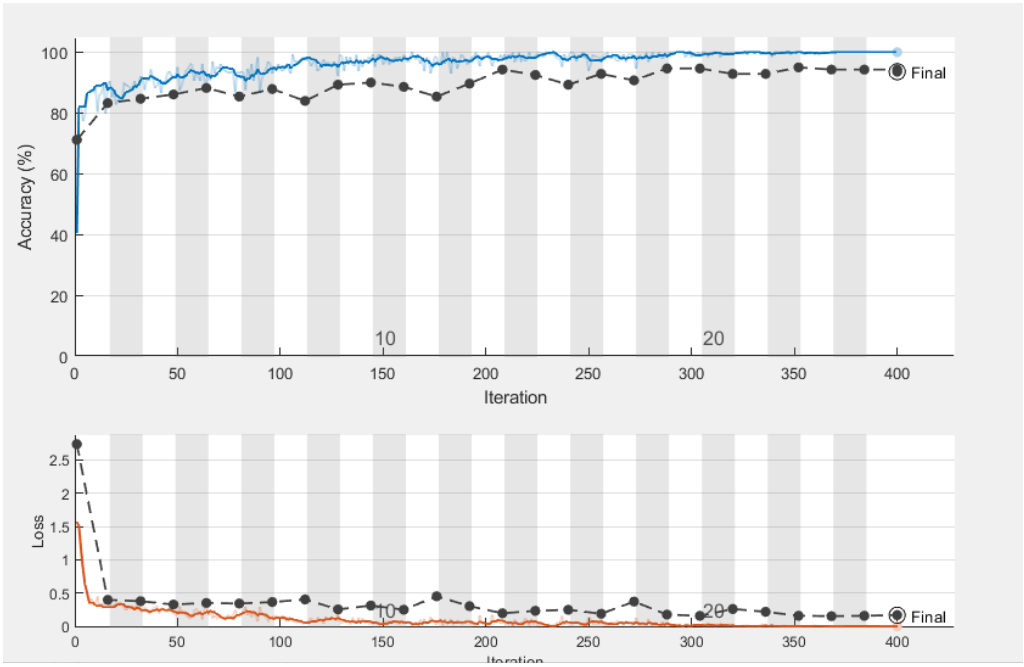


Figura 23

Experimento 2

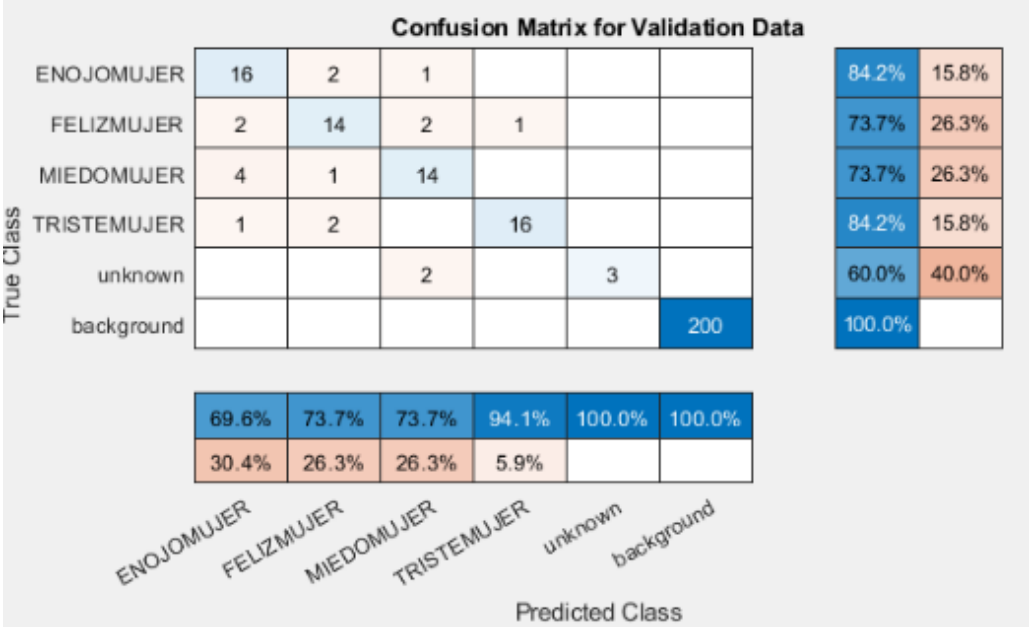


Figura 24

Experimento 3

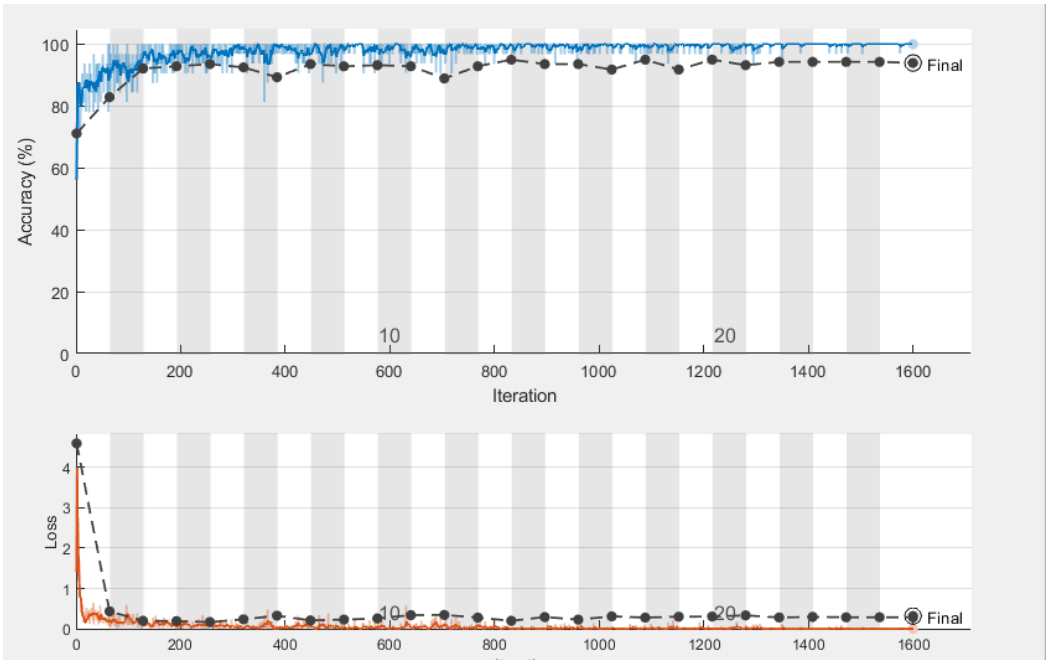


Figura 25

Experimento 3

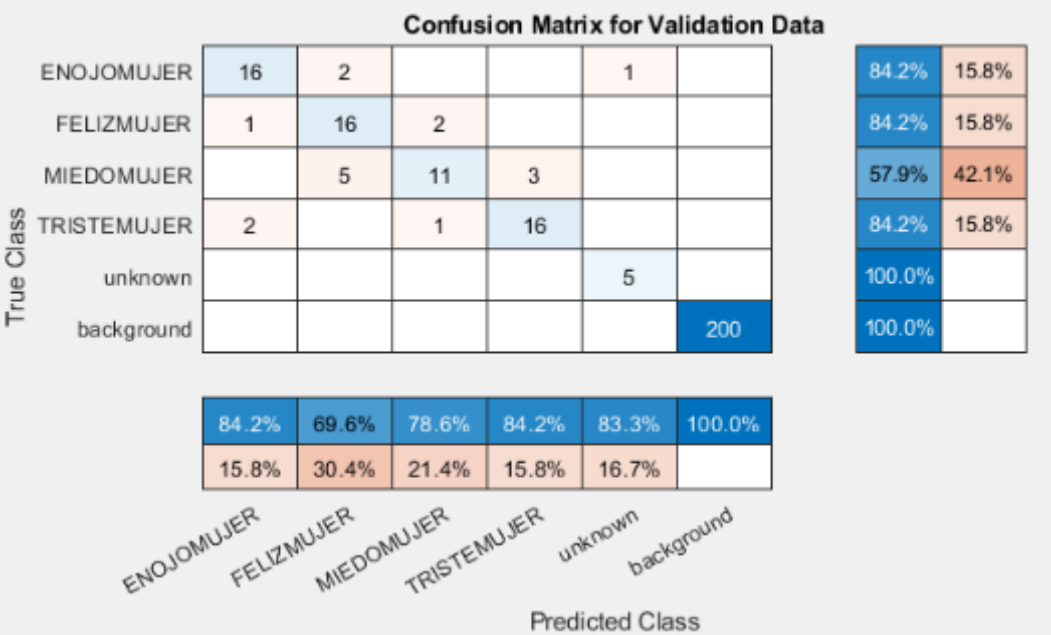


Figura 26

Experimento 4

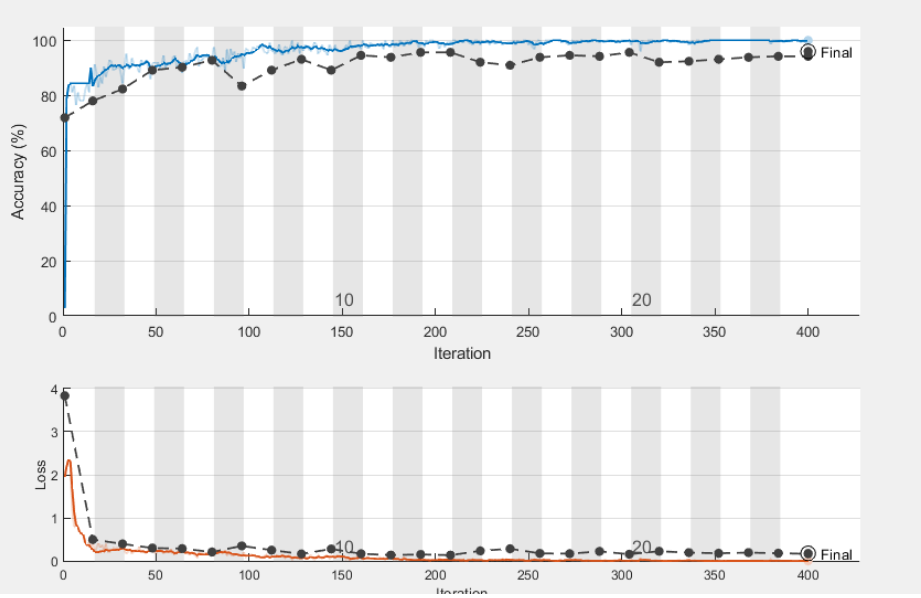


Figura 27

Experimento 4

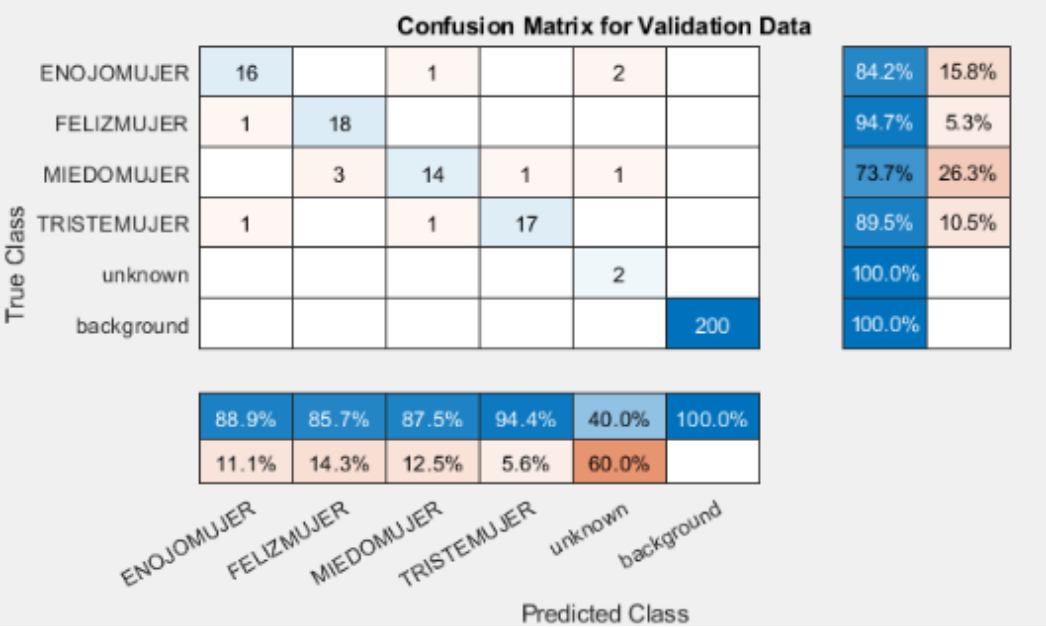


Figura 28

Experimento 5

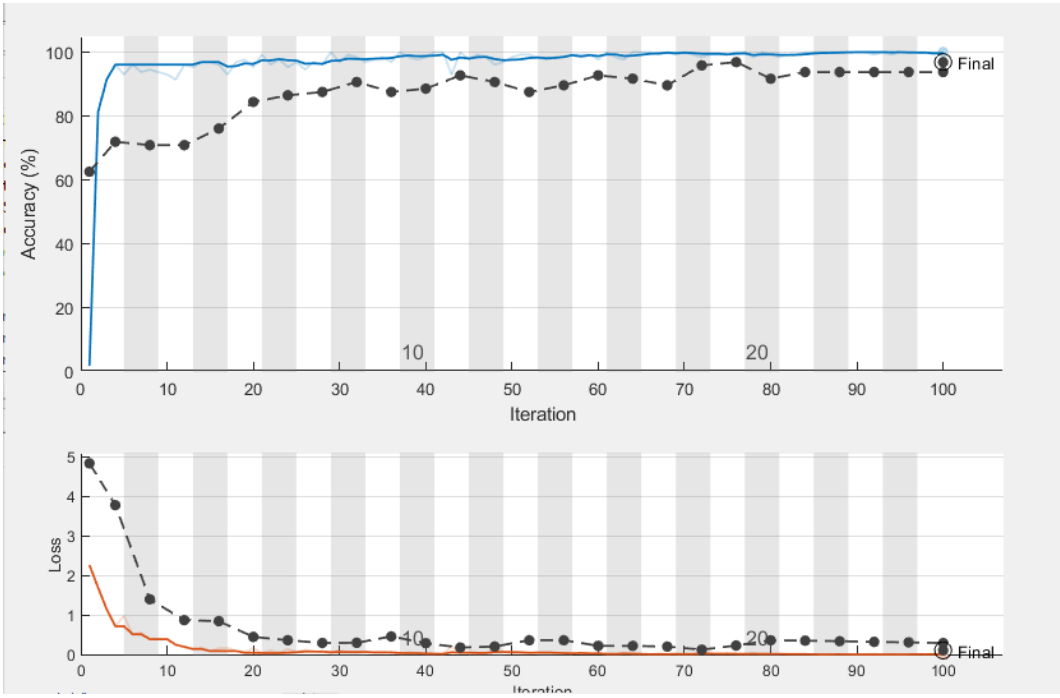
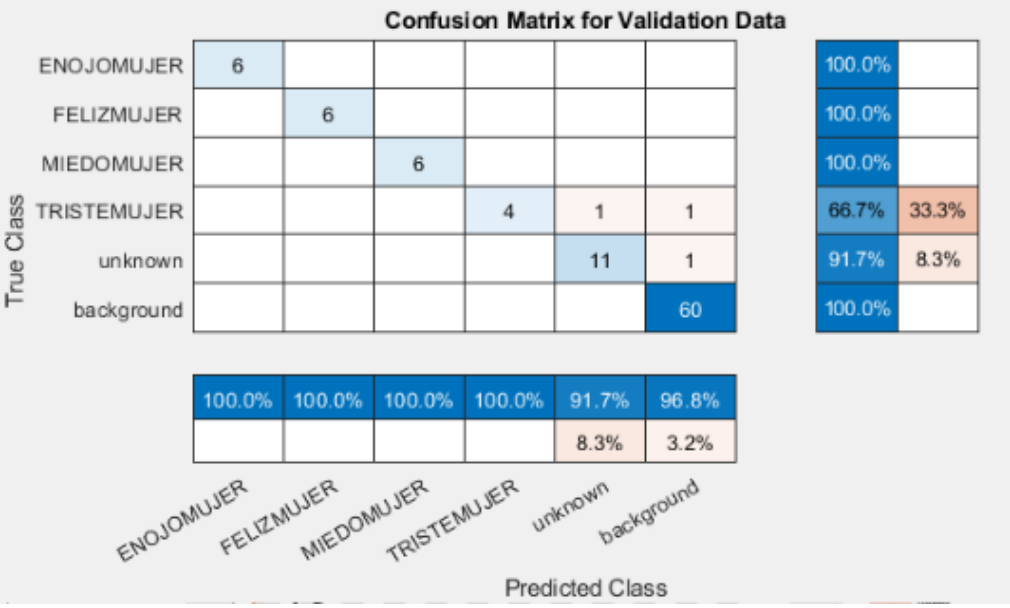


Figura 29

Experimento 5



En las Figuras 20, 22, 24, 26, 28 se puede observar el entrenamiento de la red convolucional en donde la línea azul representa la exactitud que tiene el entrenamiento de la red neuronal, la línea naranja representa las pérdidas que se tiene en la red neuronal y la línea negra presenta la validación tanto en exactitud como en pérdidas.

En las Figuras 21, 23, 25, 27, 29 se puede ver la matriz de confusión en donde se observa el porcentaje de aciertos y errores por cada emoción, con estos datos se puede obtener los resultados de exactitud, precisión, sensibilidad y especificidad.

Capítulo IV

Pruebas y resultados

Análisis de resultados

El siguiente capítulo presenta el rendimiento del sistema mediante la evaluación de ciertas características y ciertos parámetros implementados en la arquitectura de las redes convolucionales, presentando los resultados obtenidos de la base de datos propuesta en el capítulo 3.

A continuación, se pude observar los parámetros que se utiliza para obtener los diferentes resultados en cada experimento realizado.

- número total de clips de audio (N_T),
- número de clips de audio clasificados correctamente (N_E),
- verdaderos positivos (N_{VP}),
- falsos positivos (N_{FP}),
- verdaderos negativos (N_{VN}),
- falsos negativos (N_{FN}).

Con los siguientes parámetros se va a obtener los cálculos tanto de Exactitud, Precisión, Sensibilidad y Especificidad. Estos parámetros con importantes a evaluar ya que permiten obtener mayor confiabilidad en cada experimento y por supuesto evaluar cuales son los mejores resultados tanto para el clasificador de emociones de hombres como para el clasificador de emociones de mujeres.

Las fórmulas para obtener los resultados de los experimentos son los siguientes.

Exactitud (A)

$$A(\%) = \frac{N_E}{N_T} \times 100 \quad (1)$$

Precisión (P)

$$P(\%) = \frac{N_{VP}}{N_{VP} + N_{FP}} \times 100 \quad (2)$$

Sensibilidad (R)

$$R(\%) = \frac{N_{VP}}{N_{VP} + N_{FN}} \times 100 \quad (3)$$

Especificidad (S)

$$S(\%) = \frac{N_{VN}}{N_{VN} + N_{FP}} \times 100 \quad (4)$$

Ber

$$Ber = 1 - \frac{sensibilidad + especificidad}{2} \quad (5)$$

El Ber será empleado como una medida de decisión cuando sensibilidad y especificidad varíen de un experimento a otro, considerando que un menor valor de Ber el experimento tendrá mejores resultados.

Resultados del clasificador de emociones de hombres

Experimento 1

El experimento 1 mide el desempeño que tiene el clasificador con 10 capas convolucionales y un optimizador de Adam de tamaño 32. Los resultados que se obtiene de la red neuronal presentan un error de entrenamiento de 0,10 % y un error de validación de 7,69 %, los resultados del clasificador se muestran a continuación.

Tabla 1

Resultados Experimento 1

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
ENOJOHOMBRE	87,65	100	65,51	100
FELIZHOMBRE	67,90	31,6	31,6	79,03
MIEDOHOMBRE	77.77	42,11	53,3	83,33
TRISTEHOMBRE	82,71	31,57	85,7	82,43
DESCONOCIDO	92,5	100	45,45	100

Como se puede observar en la Tabla 1 los resultados obtenidos en el primer experimento son, Exactitud de 81,70 %, Precisión 61,06 %, Sensibilidad 55,71 % y una Especificidad del 88,96 %.

Experimento 2

El experimento 2 mide el desempeño que tiene el clasificador con 10 capas convolucionales y un optimizador de Adam de tamaño 128. Los resultados que se obtiene

de la red neuronal presentan un error de entrenamiento de 0,38 % y un error de validación de 12,19 %, los resultados del clasificador se muestran a continuación.

Tabla 2

Resultados Experimento 2

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
ENOJOHOMBRE	82,75	94,73	56,25	98,18
FELIZHOMBRE	73,56	42,10	40	83,58
MIEDOHOMBRE	80,45	36,84	58,33	84
TRISTEHOMBRE	86,20	42,10	88,89	85,89
DESCONOCIDO	96,55	100	78,6	100

Como se puede observar en la Tabla 2 los resultados obtenidos en el segundo experimento son, Exactitud de 83,90 %, Precisión 63,15 %, Sensibilidad 64,41 % y una Especificidad del 90,33 %.

Experimento 3

El experimento 3 mide el desempeño que tiene el clasificador con 5 capas convolucionales y un optimizador de Adam de tamaño 32. Los resultados que se obtiene de la red neuronal presentan un error de entrenamiento de 0 % y un error de validación de 10,35 %, los resultados del clasificador se muestran a continuación.

Tabla 3

Resultados Experimento 3

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
ENOJOHOMBRE	87,5	94,73	66,67	98,11
FELIZHOMBRE	76,25	36,84	50	81,81
MIEDOHOMBRE	80	57,89	57,89	86,88
TRISTEHOMBRE	90	57,89	100	88,40
DESCONOCIDO	93,7	100	44,4	100

Como se puede observar en la Tabla 3 los resultados obtenidos en el tercer experimento son, Exactitud de 85,49 %, Precisión 69,47 %, Sensibilidad 63,79 % y una Especificidad del 91,04 %.

Experimento 4

El experimento 4 mide el desempeño que tiene el clasificador con 5 capas convolucionales y un optimizador de Adam de tamaño 128. Los resultados que se obtiene de la red neuronal presentan un error de entrenamiento de 0,09 % y un error de validación de 10,63 %, los resultados del clasificador se muestran a continuación.

Tabla 4

Resultados Experimento 4

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
ENOJOHOMBRE	84,14	84,21	61,53	94,64
FELIZHOMBRE	83,78	52,63	47,61	85,24
MIEDOHOMBRE	85,36	52,63	76,92	86,95
TRISTEHOMBRE	87,80	52,63	90,90	87,32
DESCONOCIDO	93,67	100	54,5	100

Como se puede observar en la Tabla 4 los resultados obtenidos en el cuarto experimento son, Exactitud de 86,95 %, Precisión 68,42 %, Sensibilidad 66,92 % y una Especificidad del 90,83 %.

Experimento 5

El experimento 5 mide el desempeño que tiene el clasificador con la base de datos original de la Srita. Evelyn Flores (MARLEY, 2019) con 5 capas convolucionales y un optimizador de Adam de tamaño 128. Los resultados que se obtiene de la red neuronal presentan un error de entrenamiento de 0 % y un error de validación de 3,79 %, los resultados del clasificador se muestran a continuación.

Tabla 5

Resultados Experimento 5

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
ENOJOHOMBRE	100	100	100	100
FELIZHOMBRE	89,18	50	75	90,90
MIEDOHOMBRE	100	100	100	100
TRISTEHOMBRE	81,08	100	46,2	100
DESCONOCIDO	81,08	53,8	87,5	79,31

Como se puede observar en la Tabla 5 los resultados obtenidos en el quinto experimento son, Exactitud de 90,26 %, Precisión 80,76 %, Sensibilidad 81,75 % y una Especificidad del 94,04 %.

Resultados del clasificador de emociones de mujeres

Análisis del experimento 1

El experimento 1 mide el desempeño que tiene el clasificador con 10 capas convolucionales y un optimizador de Adam de tamaño 32. Los resultados que se obtiene de la red neuronal presentan un error de entrenamiento de 0% y un error de validación de 4,28 %, los resultados del clasificador se muestran a continuación.

Tabla 6

Resultados Experimento 1

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
ENOJOMUJER	92,5	94,73	78,26	98,24
FELIZMUJER	92,5	78,94	88,23	93,65
MIEDOMUJER	91,25	73,68	87,5	92,18
TRISTEMUJER	95	89,47	89,47	96,72
DESCONOCIDO	98,75	100	80	100

Como se puede observar en la Tabla 6 los resultados obtenidos en el primer experimento son, Exactitud de 94 %, Precisión 87,36 %, Sensibilidad 84,69 % y una Especificidad del 96,15 %.

Experimento 2

El experimento 2 mide el desempeño que tiene el clasificador con 10 capas convolucionales y un optimizador de Adam de tamaño 128. Los resultados que se obtiene de la red neuronal presentan un error de entrenamiento de 0 % y un error de validación de 6,40 %, los resultados del clasificador se muestran a continuación.

Tabla 7

Resultados Experimento 2

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
ENOJOMUJER	87,65	84,21	69,56	94,82
FELIZMUJER	87,65	73,68	73,68	91,93
MIEDOMUJER	87,65	73,68	73,68	91,93
TRISTEMUJER	95,06	84,21	94,11	95,31
DESCONOCIDO	97,53	60	100	97,43

Como se puede observar en la Tabla 7 los resultados obtenidos en el segundo experimento son, Exactitud de 91,10 %, Precisión 75,15 %, Sensibilidad 82,20 % y una Especificidad 94,28 %.

Experimento 3

El experimento 3 mide el desempeño que tiene el clasificador con 5 capas convolucionales y un optimizador de Adam de tamaño 32. Los resultados que se obtiene de la red neuronal presentan un error de entrenamiento de 0 % y un error de validación de 6,04 %, los resultados del clasificador se muestran a continuación.

Tabla 8

Resultados Experimento 3

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
ENOJOMUJER	92,59	84,21	84,21	95,16
FELIZMUJER	87,65	84,21	69,56	94,82
MIEDOMUJER	86,41	57,89	78,57	88,05
TRISTEMUJER	92,59	84,21	84,21	95,16
DESCONOCIDO	98,76	100	83,3	100

Como se puede observar en la Tabla 8 los resultados obtenidos en el tercer experimento son, Exactitud de 91,6 %, Precisión 82,10 %, Sensibilidad 79,97 % y una Especificidad del 94,63 %.

Experimento 4

El experimento 4 mide el desempeño que tiene el clasificador con 5 capas convolucionales y un optimizador de Adam de tamaño 128. Los resultados que se obtiene de la red neuronal presentan un error de entrenamiento de 0 % y un error de validación de 3,94 %, los resultados del clasificador se muestran a continuación.

Tabla 9

Resultados Experimento 4

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
ENOJOMUJER	93,58	84,21	88,88	95
FELIZMUJER	94,87	94,73	85,71	98,24
MIEDOMUJER	91,02	73,68	87,5	91,93
TRISTEMUJER	96,15	89,47	94,44	96,67
DESCONOCIDO	96,15	100	40	100

Como se puede observar en la Tabla 9 los resultados obtenidos en el cuarto experimento son, Exactitud de 94,35 %, Precisión 88,41 %, Sensibilidad 79,31 % y una Especificidad del 96,36 %.

Experimento 5

El experimento 5 mide el desempeño que tiene el clasificador con la base de datos original de la Srita. Evelyn Flores (MARLEY, 2019) con 5 capas convolucionales y un optimizador de Adam de tamaño 128. Los resultados que se obtiene de la red neuronal presentan un error de entrenamiento de 0 % y un error de validación de 3,12 %, los resultados del clasificador se muestran a continuación.

Tabla 10

Resultados Experimento 5

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)
ENOJOHOMBRE	100	100	100	100
FELIZHOMBRE	100	100	100	100
MIEDOHOMBRE	100	100	100	100
TRISTEHOMBRE	94,28	66,67	100	93,54
DESCONOCIDO	97,05	91,7	91,7	95,65

Como se puede observar en la Tabla 10 los resultados obtenidos en el quinto experimento son, exactitud de 98,26 %, precisión 91,67 %, sensibilidad 98,34 % y una especificidad del 97,83 %.

Resultado total del clasificador de emociones de hombres

A continuación, en la Tabla 11 se puede observar los resultados totales que se obtuvieron en los diferentes experimentos del clasificador de emociones de hombres.

Tabla 11

Resultados del Clasificador de Emociones de Hombres

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)	Ber
Experimento 1	81,70 %	61,01 %	55,71 %	88,96 %	0,2767
Experimento 2	83,90 %	63,15 %	64,41 %	90,33 %	0,2263
Experimento 3	85,49 %	69,47 %	63, 79 %	91,04 %	0,2258
Experimento 4	**86,95 %**	**68,42 %**	**66,92 %**	**90,83 %**	**0,2112**
Experimento 5	90,26 %	80,76 %	81, 75 %	94,04 %	0,1211

En la Tabla 11 se puede observar que el mejor experimento para el clasificador de hombres en los primeros cuatro experimentos con la base de datos de 80 audios por cada emoción es el Experimento 4 con un cálculo de Ber de 0,2112. Mientras que el Experimento 5 con una base de datos de 13 audios por cada emoción presenta mejores resultados ante todos los experimentos con un cálculo de Ber de 0,1211

Resultado total del clasificador de emociones de mujeres

A continuación, en la Tabla 12 se puede observar los resultados totales que se obtuvieron en los diferentes experimentos del clasificador de emociones de mujeres.

Tabla 12

Resultados del Clasificador de Emociones de Mujeres

	Exactitud (%)	Precisión (%)	Sensibilidad (%)	Especificidad (%)	Ber
Experimento 1	**94 %**	**87,36 %**	**84,69 %**	**96,15 %**	**0,0958**
Experimento 2	91,10 %	75,15 %	82,20 %	94,28 %	0,1176
Experimento 3	91,6 %	82,10 %	79,97 %	94,63 %	0,1270
Experimento 4	94,35 %	88,41 %	79,31 %	96,36 %	0,1216
Experimento 5	98,26 %	91,67 %	98,34 %	97,83 %	0,0192

En la Tabla 12 se puede observar que el mejor experimento para el clasificador de mujeres en los primeros cuatro experimentos con la base de datos de 80 audios por cada emoción es el Experimento 1 con un cálculo de Ber de 0,0958. Mientras que el Experimento 5 con una base de datos de 13 audios por cada emoción presenta mejores resultados ante todos los experimentos con un cálculo de Ber de 0,0192.

Capítulo v

Conclusiones y Recomendaciones

- De la Tabla 11 se concluyó que el mejor método para el clasificador de emociones de hombres es utilizando un optimizador de Adam de 128 y una red convolucional de 5 capas.
- De la Tabla 12 se concluyó que el mejor método para el clasificador de emociones de mujeres es utilizando un optimizador de Adam de 32 y una red convolucional de 10 capas.
- De las Tablas 11 y Tabla 12, se puede verificar que el experimento 5 tiene mejores resultados con relación a los demás experimentos, estos resultados se obtuvo con una base de datos total de 26 audios los cuales se reparte equitativamente entre hombres y mujeres, la base de datos en mención esta mejor elaborada ya que tiene audios con mejores características para un entrenamiento y validación, estos audios permite que la red neuronal convolucional refleje valores de exactitud, precisión, sensibilidad y especificidad de hasta un 100 %.
- Los parámetros que se emplean para el entrenamiento de la red neuronal tales como el cambio y la variación en la capa convolucional y en el optimizador de Adam, son parámetros que obtuvo mejores resultados a los diferentes experimentos realizados, con esto se pudo obtener valores mínimos tanto en error de entrenamiento como error de validación.

.

- El problema más complejo de abordar en el presente trabajo fue la base de datos que se encontró en uso libre (RAVDESS), esta base de datos presentaba diferencia en algunas características de algunos audios tales como la frecuencia

de muestro, tamaño y velocidad en bits, por lo que se recomienda buscar convenios para tener acceso a bases de datos mejor elaboradas.

- A pesar de que el proceso de clasificación de cada emoción resulte ser un trabajo laborioso es recomendable revisar los formatos de cada archivo de audio y las etiquetas, debido a que al estar los archivos ya etiquetados en la base de datos se determinó que no se encontraban con su etiqueta correcta.

Capítulo VI

Líneas de trabajos futuros

Trabajos Futuros

- Se puede realizar una comparación y observar los resultados que se obtiene entre el uso de la escala Mel y el uso de la escala Bark y verificar también la diferencia entre los espectrogramas que se obtendrán.
- Aumentar la base de datos con otras emociones, tales como disgusto, sorpresa, calma, temeroso.
- Se puede realizar un sistema más autónomo con el uso de micrófono para el ingreso de datos.

Bibliografía

Anton Batliner, B. S. (01 de 2011). *The Automatic Recognition of Emotions.* https://www.researchgate.net/publication/224929599_The_Automatic_Recognition_of_Emotions_in_Speech

Arteaga, G. (2015). *Aplicación del aprendizaje profundo ("deep learning") al procesamiento de señales digitales.* https://red.uao.edu.co/bitstream/10614/7975/1/T05978.pdf

Barclay, T. (16 de 07 de 2019). *Larynx*. Innerbody: https://www.innerbody.com/anatomy/respiratory/head-neck/larynx

Bericat, E. (2012). *Emotions.* Emotions: http://www.sagepub.net/isa/resources/pdf/Emociones.pdf

Berkehan Akça, M., & Oğuz, K. (Enero de 2020). *Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers.* Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers: https://www.sciencedirect.com/science/article/pii/S0167639319302262#bib0139

Celdrán Baños, J., & Ferrándiz García, C. (2012). *Reconocimiento de emociones en niños de Educación Primaria.* Eficacia de un programa educativo para reconocer emociones: http://repositorio.ual.es/bitstream/handle/10835/1908/Art_28_718.pdf?sequence=1

Chen, C. J. (2016). *Elements of Human Voice.* World Scientific Publishing Co. Pte. Ltd.: http://www.columbia.edu/~jcc2161/documents/HumanVoice.pdf

Chóliz, M. (1995). *La expresión de las emociones en la obra de Darwin.* Valencia: Promolibro.

Hernández , R., Pérez, E., Orozco , D., Sánchez, L., & Hidalgo, M. (Marzo de 2018). *Deep Learning. Una revisión .* https://www.researchgate.net/publication/323858502_Deep_Learning_Una_revision

INSTRUMENT, T. (2018). *Introduction to Deep Learning.* Ejemplo de una Red neuronal convolucional (CNN): https://training.ti.com/sites/default/files/docs/introduction-to-deep-learning.pdf

Kingma, D. P., & Ba, J. (22 de Diciembre de 2014). *Adam: A Method for Stochastic Optimization*. Adam: A Method for Stochastic Optimization: https://arxiv.org/abs/1412.6980

LeCun, Bengio, Y., & G. Hinton. (2015). Deep learning. *Nature*, 436-444.

Livingstone, S. R., & Russo, F. A. (5 de Abril de 2018). *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. https://zenodo.org/record/1188976#.X6vtrGhKjIW

MathWorks. (2021). *batchNormalizationLayer*. https://www.mathworks.com/help/deeplearning/ref/nnet.cnn.layer.batchnormalizationlayer.html

MathWorks. (2021). *convolution2dLayer*. convolution2dLayer: https://www.mathworks.com/help/deeplearning/ref/nnet.cnn.layer.convolution2dlayer.html

MathWorks. (2021). *fullyConnectedLayer*. https://www.mathworks.com/help/deeplearning/ref/nnet.cnn.layer.fullyconnectedlayer.html

MathWorks. (2021). *maxPooling2dLayer*. https://www.mathworks.com/help/deeplearning/ref/nnet.cnn.layer.maxpooling2dlayer.html

MathWorks. (2021). *reluLayer*. https://de.mathworks.com/help/deeplearning/ref/nnet.cnn.layer.relulayer.html;jsessionid=b44161541ee061004f4d798b28b9

Mirsamadi, S., Barsoum, E., & Cha, Z. (13 de Marzo de 2017). *Automatic.* Autdallas: https://personal.utdallas.edu/~mirsamadi/files/mirsamadi17a.pdf

Missinglink.ai. (2018). *Fully Connected Layers in Convolutional Neural Networks: The Complete Guide*.

Morzaria, H. (04 de 12 de 2019). *The Different Types of Emotions and How They Impact Human Behavior*. Business 2 community: https://www.business2community.com/workplace-culture/the-different-types-of-emotions-and-how-they-impact-human-behavior-02263872

Rouse, M. (April de 2018). *Convolutional neural network.* SearchEnterpriseAI: https://searchenterpriseai.techtarget.com/definition/convolutional-neural-network

Savyakhosla. (29 de Julio de 2021). *CNN | Introduction to Pooling Layer*. GeeksforGeeks: https://www.geeksforgeeks.org/cnn-introduction-to-pooling-layer/

Schuller, B. W. (05 de 2018). *Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends*. Communications of the ACM: https://cacm.acm.org/magazines/2018/5/227191-speech-emotion-recognition/fulltext

SHARMA, V. (3 de Mayo de 2020). *Role of Convolutional Layer in Convolutional Neural Networks*. Vinodsblog: https://vinodsblog.com/2020/05/03/role-of-convolutional-layer-in-cnn/

The voice fundation. (2017). *The voice fundation*. Health science: https://voicefoundation.org/health-science/voice-disorders/anatomy-physiology-of-voice-production/understanding-voice-production/

Torres, D. B. (2007). *Anatomia funcional de la voz*. Diposit: http://diposit.ub.edu/dspace/bitstream/2445/43135/1/anatomia-funcional-voz.pdf

Ververidis, D., & Kotropoulos, C. (2006). *Emotional speech recognition: Resources, features, and methods.* Emotional speech recognition: Resources, features, and methods: https://www.sciencedirect.com/science/article/abs/pii/S0167639306000422

Zhang, Z. (2016). *Mechanics of human voice production and control*. JASA: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5412481/