EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

INSTITUTO TECNOLÓGICO DE CULIACÁN

SISTEMA DE IDENTIFICACIÓN POR VOZ MEDIANTE EL USO DE REDES PROFUNDAS

TESIS

PRESENTADA ANTE EL DEPARTAMENTO ACADÉMICO DE ESTUDIOS DE POSGRADO DEL INSTITUTO TECNOLÓGICO DE CULIACÁN EN CUMPLIMIENTO PARCIAL DE LOS REQUISITOS PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

POR:

ING. JOSÉ LUIS MEDINA JIMÉNEZ
INGENIERO EN MECATRÓNICA

DIRECTOR DE TESIS:

MC. GLORIA EKATERINE PERALTA PEÑUÑURI

CULIACÁN, SINALOA AGOSTO 2022

Dedicatoria

A mis **padres**.

Agradecimientos

Agradezco completa e infinitamente a mis padres **Yolanda Jiménez Vizcarra** y **Efraín Medina Valentón** por ser siempre mis pilares y ejemplos a seguir, por jamás soltarme de la mano en mi crecimiento y apoyarme en todo cada día, por enseñarme a que con esfuerzo se puede lograr todo lo que se proponga.

A mis **abuelos** Efraín Medina Alcaraz y María Valentón Beltrán por siempre ser un apoyo en mi vida y ser unos segundos padres para mí, por estar siempre en las buenas y malas apoyándome.

A mis **hermanos** Brayan Medina y Maria Medina por ser mi motivación de superarme día a día y ser un ejemplo a seguir para ellos.

A mi **asesores** de tesis MC. Gloria Peralta y Dr. Héctor Rodríguez por asesorarme y guiarme tanto profesional como personalmente para lograr concluir este trabajo realizado con esfuerzo y dedicación.

A todos los **Profesores** de la maestría en Ciencias de la Computación por trasmitirnos sus conocimientos que me ayudaron en mi desarrollo profesional.

A mis **compañeros** de maestría por ser buenos compañeros y amigos, apoyándonos siempre en todo momento, siendo grandes personas.

Al **Instituto Tecnológico de Culiacán** por ser parte de mi crecimiento en mi formación profesional, obteniendo el grado de Ing. Mecatrónico y ahora en maestría.

Al **Consejo Nacional de Ciencia y Tecnología** por ayudarme con el apoyo económico de mis estudios de maestría.

Declaración de Autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

José Luis Medina Jiménez. Culiacán, Sinaloa, México, 2022.

Resumen

En la actualidad los métodos tradicionales de autentificación, como los pines de seguridad o el usuario y contraseña, se han convertido un problema debido a que son fáciles de olvidar y no proporcionan la seguridad de resguardar los datos de manera confiable.

La seguridad biométrica se ha explorado cada vez más debido a que los rasgos biométricos son únicos como método de autentificación. Además, gracias al avance de la inteligencia artificial y de los métodos de extracción de características, la seguridad biométrica cada vez se vuelve más precisa.

En este trabajo, se realizó un sistema de identificación por voz utilizando una red neuronal convolucional llamada VGGVox. Además, se utilizó como método de extracción de características los espectrogramas de los audios, con la finalidad de obtener una representación del audio del dominio del tiempo en el dominio de las frecuencias para posteriormente realizar la identificación.

Por otra parte, se desarrolló una base de datos con distintos alumnos del Instituto Tecnológico de Culiacán, en el cual participaron 128 personas. La base de datos contiene tres audios diciendo la misma frase con la que se registraron y otros tres audios diciendo frases distintas de una lista dada. En total se recopilaron 768 audios como base de datos.

Por ultimo, la base de datos fue utilizada dentro del sistema de identificación por voz, con la finalidad de realizzar un análisis de distintos umbrales de aceptación para aumentar la precisión del sistema, logrando llegar a obtener un 93 %

Palabras Clave

- Aprendizaje máquina
- Aprendizaje profundo
- Autentificación de voz
- Biometría
- Identificación de voz
- Inteligencia artificial
- Reconocimiento de voz

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El desarrollo de la tecnología trajo consigo el crecimiento de la era digital, de esta manera, cada vez era más requerida la autentificación de personas para tener control y seguridad en la información personal. El robo de identidad y datos personales ha crecido de la misma manera que la era digital, por lo cual, la falsificación y suplantación ha afectado a millones. Según datos del banco de México (Banxico) el país se encuentra en octavo lugar en el mundo en delito de robo de identidad (Banxico, 2021).

Actualmente, los métodos tradicionales para autentificación, como lo es el usuario y contraseña, ya no son suficientes debido a que son fáciles de olvidar o en el peor de los casos ser robados para la suplantación de identidad. Es por ello que los rasgos biométricos cada vez son de gran importancia, ya que son características que definen a las personas como únicas unas de otras (Gayathri et al., 2019).

Por otro lado, los pines de seguridad son métodos de autentificación en los cuales es necesario contar siempre con ellos a la mano un dispositivo que los genere, donde estos corren el riesgo de que puedan ser clonados y exponer la información que se resguarda a través de estos métodos.

Los rasgos biométricos son considerados como las características únicas que tiene cada individuo, las cuales pueden ser tanto físicas como de comportamiento, donde las físicas pueden ser huella dactilar, iris, geometría de la mano, cara, así como las de comportamiento se conforman por la firma, manera de caminar, escritura, etc.

El reconocimiento y verificación biométrica son aspectos importantes de investigación, ya que hoy en día la contraseña es considerada como un método tradicional y muy vulnerable, de

tal manera que es reemplazada por los rasgos biométricos en distintos sectores de aplicaciones. Para esto es importante tomar en cuenta que la finalidad de utilizar los datos biométricos como sistema de autentificación es aumentar la seguridad y privacidad del usuario, ya que los datos biométricos son únicos. (Gayathri et al., 2019)

Las técnicas de la inteligencia artificial han sido de gran utilidad para el desarrollo e implementación del reconocimiento biométrico, y más aún el aprendizaje profundo ha tomado gran importancia para el desarrollo de sistemas biométricos, siendo algunas de las arquitecturas de aprendizaje profundo más usadas por la comunidad de visión computacional, las cuales incluyen, redes convolucionales, redes recurrentes, en especial las llamadas memoria a corto plazo o por sus siglas en inglés LSTM, autocodificadores y las redes generativas adversarias (GANs) (Minaee et al., 2019).

El proyecto de investigación, desarrollado, está enfocado en la implementación de un sistema de identificación de voz. La identificación de voz es una rama del reconocimiento de voz, el cual se centra en procesar una muestra de voz desconocida y compararla con una base de datos de personas establecidas. La voz desconocida se identifica como la que mejor se adapte, de esta manera la entrada para la identificación de voz es una voz desconocida y la salida es el nombre o la identificación del usuario (Tandel et al., 2020).

1.1. Definición del problema

El aumentar la seguridad y privacidad de los usuarios en los sistemas al momento de autentificarse ha hecho que se busquen alternativas a los distintos métodos tradicionales, tales como usuario y contraseña, llaves de acceso y pines de seguridad. Debido a que estos métodos de autentificación se pueden perder, olvidar o inclusive ser robados para usos ilícitos.

Actualmente, el uso desmedido del internet ha ocasionado que se tengan múltiples cuentas en cualquier sitio, logrando que el manejo de distintas contraseñas para todas las cuentas sea casi imposible. Por otro lado, en áreas de seguridad como el banco, es difícil solo mantener zonas seguras con manejo de pines de seguridad para el acceso de personal autorizado, ya que estos pines fácilmente puede introducirlos cualquier persona sin tener la certeza de que pertenezca a la institución.

Los datos biométricos juegan un papel importante hoy en día, pasando a ser un sustituto idóneo de los métodos de autentificación tradicionales por la sencilla y gran razón de ser características únicas en cada individuo, además de contar con múltiples rasgos biométricos para validación de usuarios.

Existen casos en el área forense donde hay grabaciones de audio y no se cuenta con algún video o los videos son de mala calidad, de tal manera que no se puede identificar a la persona. Es aquí donde entran los sistemas de reconocimiento de voz, con los cuales se vuelven indispensables si la única manera de identificar a alguien es mediante grabaciones de audio.

Los proveedores de asistencia telefónica no pueden tener la certeza de quien está del otro lado del teléfono, debido a que la única forma de validación es pidiendo datos personales como fecha de nacimiento, nombre, y/o un ID. Siendo un área de oportunidad para la implementación de autentificación por voz para explorar alternativas al momento de autentificar alguien mediante asistencia telefónica.

La voz es uno de los rasgos menos explorados en comparación con el reconocimiento mediante el rostro, iris y huella dactilar. Aunque es un rasgo único y con el que se puede acceder más fácilmente, ya que todo mundo cuenta con un micrófono en cualquier dispositivo. También resulta ser un rasgo complejo de estudiar esto debido a que la voz depende de distintos factores, tanto físicos como de comportamiento.

Mediante el timbre de voz se detecta las frecuencias, las cuales son una de las características con las que cuenta un audio. Estas frecuencias suelen variar dependiendo de las emociones, articulación de palabras, la distancia del micrófono, ruidos ambientales, la diferencia de edad o inclusive infecciones de la garganta, suelen ser algunas de las problemáticas y retos que representa el estudio del reconocimiento por voz.

1.2. Hipótesis

La implementación de las técnicas de aprendizaje profundo permitirán desarrollar un sistema biométrico de identificación por voz sin utilizar una frase específica, con el fin de emplearlo como método de autentificación con una precisión superior al 90 %.

1.3. Objetivos

1.3.1. Objetivo general

Diseñar e implementar un sistema biométrico de identificación por voz utilizando técnicas de aprendizaje profundo para explorar la confiabilidad del sistema al momento de auntetificar y desarrollar una base de datos de voces con frases distintas para la validación del sistema.

1.3.2. Objetivos específicos

- Seleccionar técnicas de aprendizaje profundo para el reconocimiento de voz.
- Diseñar y crear una base de datos de voces con distintas frases para pruebas y validación del sistema.
- Implementar la técnica de aprendizaje profundo seleccionada para el desarrollo de un sistema de identificación por voz como método de autentificación.
- Validar el sistema de identificación de voz mediante el uso de la base de datos de voces diseñado para analizar y validar los resultados obtenidos.

1.4. Justificación

El área de integración de la biometría en sistemas como método de autentificación tiene gran auge y se han vuelto un gran reto, es por eso, que aún se siguen realizando diversas investigaciones de los distintos rasgos biométricos que distinguen a cada individuo.

En la actualidad diversas áreas han implementado los sistemas biométricos, tales como sistemas bancarios para acceso de personal a áreas restringidas y usuarios del banco para mantener seguras las cuentas bancarias. En dispositivos móviles como método de ingreso a las funciones, usando comúnmente la huella dactilar o el reconocimiento facial, siendo estos los más explotados actualmente. Además, el desarrollar una herramienta capaz de identificar una persona mediante la voz, abre las posibilidades de aplicarse a las ciencias criminalísticas para obtener pruebas que sean válidas para un juicio.

Los sistemas unimodales biométricos, definidos como, los sistemas de un solo rasgo biométrico como autentificador, puede parecer tener un gran impacto a la hora de incrementar la seguridad y privacidad de usuario, pero la implementación multimodal biométrica puede ser mayormente atractiva. En otras palabras, el uso de múltiples rasgos biométricos para autentificar a una persona y dar acceso resulta ser mayormente beneficioso si se requiere un sistema robusto y altamente seguro para la identificación de usuarios.

Por otro lado, la implementación de los sistemas de autentificación por voz dentro de áreas de asistencia telefónica como en bancos para validar a una persona quien dice ser y mediante esto, tomar una decisión si la llamada compromete o no los datos del propietario, además que facilita el proceso de ingreso a los usuarios, evitando un cuestionario de datos para ingresar. Cabe aclarar que este tipo de validación en específico es necesario contar con sistemas especializados en detección de falsificación (AntiSpoofing), el cual es un tema de estudio diferente al de este proyecto.

Por último, el uso de la voz como control de acceso a una vivienda o áreas restringidas como edificios o secciones de edificios, permite y facilita que con solo una grabación o frase sea validada una persona para ingresar a un lugar, siempre y cuando cuente con los permisos necesarios.

Un sistema de acceso por voz, es capaz de cumplir con las medidas sanitarias que se implementaron a partir de la pandemia por el virus COVID-19, ya que es un tipo de sistema no invasivo en una persona, como lo puede ser el ingresar mediante usuario y contraseña o el usar la huella dactilar, evitando así el contacto físico y propagación de cualquier virus contagioso al contacto. [1]

Con el fin de encontrar nuevas herramientas que puedan crear sistemas más robustos al complementar diversos rasgos biométricos como método de acceso, los sistemas de reconocimiento por voz son una buena solución para crear métodos de autentificación más seguros y confiables que no solo dependan de un rasgo biométrico, reduciendo lo más posible las probabilidades de suplantación de identidad que se vive hoy en día.

[1] https://coronavirus.gob.mx/medidas-de-seguridad-sanitaria/

1.5. Estructura de la tesis

El trabajo actual está organizado por 6 capítulos, donde el primer capítulo es el actual de Introducción.

- Capítulo 2 - Marco Teórico: El capítulo incluye los fundamentos en los que se basa el trabajo sobre inteligencia artificial, los sistemas de aprendizaje-máquina, la biometría, sistemas biométricos, de esta forma se obtiene la información que sustente el desarrollo del proyecto.

- Capítulo 3 - Estado del Arte: Muestra los trabajos más recientes en la actualidad enfocados en el tema de sistemas biométricos implementados en el reconocimiento de la voz y conocer como fueron abordados estos proyectos.

- Capítulo 4 - Metodología: Se muestra el desarrollo y pasos a seguir para la implementación del sistema biométrico, es decir, el proceso que se llevó a cabo para obtener como resultado la etapa final del proyecto, así como las herramientas implementadas para llegar a estos resultados.

- Capítulo 5 - Análisis de Resultados: Se presenta todo lo obtenido con el sistema desarrollado, tal como gráficas, pruebas en la implementación del modelo utilizado en la metodología y el análisis de los tiempos de entrenamiento y precisión del sistema.

- Capítulo 6 - Conclusiones: Se muestra un resumen de la forma en que se llegó a los resultados y el aporte del proyecto al área de seguridad en métodos de autentificación. Además, se toman en cuenta trabajos a futuros para mejorar el sistema.

Capítulo 2

Marco teórico

Para el desarrollo del proyecto es necesario contar con bases importantes que sustenten este. A continuación se detalla el capítulo que tiene como propósito plantear aquellos conceptos que son de suma importancia, donde se tratará temas, desde lo más general hasta lo más especifico. Dentro de los conceptos claves se tiene a la inteligencia artificial, redes neuronales, aprendizaje-máquina, aprendizaje profundo e hiperparámetros. Por último, se abordarán los conceptos del tema de investigación en particular, es decir, conceptos enfocados a los sistemas de identificación/verificación de voz.

2.1. Inteligencia Artificial

Para entender lo que es la inteligencia artificial es necesario conocer el término o a que se refiere la palabra inteligencia, según (Essential English Dictionary, Collins, London, 2008) la habilidad de razonar las cosas en lugar de hacerlas por simple instinto o de manera autómata. Ahora bien, según (Negnevitsky, 2005) se define como aquella ciencia donde la máquina hace cosas que requieren razonamiento hecho por el humano. Es decir, donde se tiene la capacidad de tomas decisiones con base en los datos obtenidos.

Otra definición, según (Omil, 2019) la inteligencia artificial es una habilidad de un ordenador o conjunto de ordenadores, el cual realiza las tareas que haría un humano inteligente.

Historia

En 1943, Warren McCulloch y Walter Pitts (Negnevitsky, 2005) presentaron el primer trabajo reconocido del área de la Inteligencia Artificial, ellos contribuyeron con un modelo de neuronas del cerebro, considerado como la primera mayor contribución para la Inteligencia Artificial (IA). Estos mismos autores propusieron un sistema de redes neuronales el cual cada neurona utiliza un estado binario, es decir, un sistema tipo On-Off, demostrando que este modelo es igual a la máquina de Turing, a su vez demostraron que esta estructura neuronal podría aprender. En 1958, McCarthy (Negnevitsky, 2005) presentó un programa denominado "*Advice Taker*" en el artículo llamado "*Programs with common sense*", este programa se enfocaba en encontrar soluciones a problemas generales. Se consideró como el primer sistema que representaba el conocimiento y el razonamiento. Durante el periodo de 1961 a 1972, existió un proyecto de grandes expectativas llamado General Problem Solver (GPS) propuesto por Allen Newel y Herbert Simon, este tenía como propósito el resolver problemas mediante los métodos de los humanos. Este sistema se basaba en el método conocido como análisis de medio-fin. Por otro lado, en 1965 Lotfi Zadeh (Negnevitsky, 2005), publicó un artículo considerado como los principios de la teoría de conjuntos difusos, este artículo tiene por nombre como "*Fuzzy sets*" el cual sirvió para años más adelante ser la base para el desarrollo de máquinas y sistemas inteligentes. Para los años 70 los investigadores se percataron que el hecho de querer resolver problemas generales como lo haría el humano no era la mejor perspectiva, dando pie a esto a que el dominio del problema de las máquinas inteligentes fueran más enfocados a resolver problemas específicos y no de carácter general. Un ejemplo importante es el sistema DENDRAL desarrollado Buchanan y otros en 1969 el cual tiene como tarea hacer análisis químicos, este sistema fue desarrollado para hacer análisis molecular del suelo en una expedición a Marte. De este modo es como nacen los sistemas expertos, ya que para poder desarrollar un sistema enfocado a una tarea específica es necesario de que al menos una persona entre las encargadas a desarrollar el sistema tenga el conocimiento suficiente para ser resolutivo en el área. Más adelante a mediados de los 80 los sistemas expertos no fueron suficientes, es por ello por lo que las redes neuronales volvieron al campo de investigación, ya que su desarrollo estuvo detenido debido a que no se contaba con la capacidad

computacional para llevarlas a cabo. Una de las mayores aportaciones durante esta década fue el desarrollo de la red Hopfield hecha por Hopfield en 1982, el cual es una red neuronal con retroalimentación. en 1986 Rumelhart y McClelland (Negnevitsky, 2005) mejoraron el algoritmo de aprendizaje por retroalimentación basándose en el trabajo por Bryson and Ho en 1969 (Negnevitsky, 2005). A su vez, el aprendizaje por retroalimentación fue descubierta por Parker y LeCun en 1987 y 1988 respectivamente (Negnevitsky, 2005), y a partir de esa época esta técnica se volvió muy utilizada para el entrenamiento de las redes multicapa. Por otro lado, en el mismo año de 1988 dos investigadores Broomhead y Lowe (Negnevitsky, 2005) desarrollaron un procedimiento para realizar redes de alimentación en capas de base radial, que se volvería una alternativa a las redes multicapa.

2.2. Redes Neuronales

Una red neuronal está conformada por una capa de entrada de neuronas, de una a más capas ocultas de neuronas y por último una capa de salida, todas estas capas están conectadas entre sí con un peso asociado a cada conexión como se muestra en la Figura 2.1. Cada capa o neurona requiere los datos de entrada de la anterior, para esto es necesario una función de activación, la cual es la que se encarga de entregar una salida con estos datos. La Ecuación 2.1 define la parte matemática de la red neuronal artificial.

$$h_i = \sigma\left(\sum_{j=1}^{N} V_{ij}X_j + T_i^{hid}\right) \tag{2.1}$$

h_i es la salida de la neurona donde "i" se refiere a la capa que va iterando, $\sigma()$ esta parte del sigma hace referencia a la función de activación que se mencionó anteriormente, cuál es su propósito, donde dentro de ella tenemos N que es el número de neuronas de entrada, tenemos v_ij la cual son los pesos y x_j son los valores de entrada y por último está T_i^hid el cual es el umbral de las neuronas en la capa oculta. (Wang, 2003)

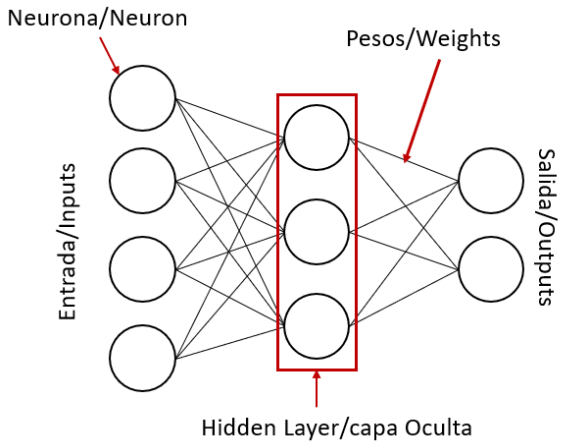

Figura 2.1: Arquitectura de una Red Neuronal

Una red neuronal artificial son modelos electrónicos basados en una neuronal del cerebro. La idea de replicar una neurona de forma artificial es de hacerla aprender de la misma manera como lo haría una neurona biológica, es decir, mediante la experiencia. Las redes neuronales están conformadas por unidades de procesamiento sencillas que conforme se van conectando unas con otras generan una red muy bien estructurada. Ahora bien, cada nodo o las unidades de procesamiento que cuenta las redes es la simplificación de una neurona biológica, la cual dispara señales de entrada a las unidades con las que está comunicada (Dongare et al., 2012).

Como ya se mencionó anteriormente las RNA utilizan el paso de información de unas con otras, para esto es necesario una función de activación, existen distintas funciones de activación, cada una se caracteriza por su ecuación o su condición para dar paso a la información.

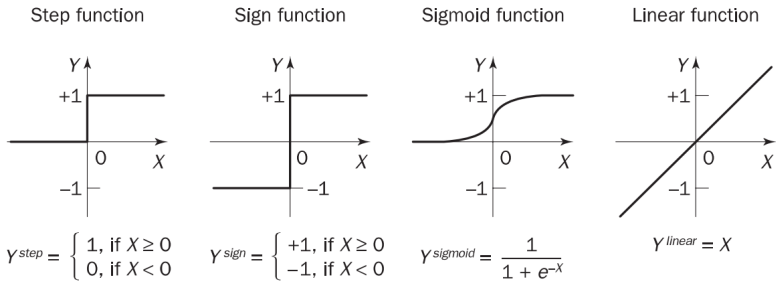

Figura 2.2: Funciones de activación((Negnevitsky, 2005))

En la Figura 2.2 se observan que las primeras dos funciones son similares, y tienen un nombre especial, el cual es "*hard-limit functions*" esto debido a que si se cumple con el um-

bral la función forza a que la salida sea 1 de otra manera sea 0. La función sigmoide normaliza los valores de entrada en valores entre 0 y 1. Esta función es normalmente utilizada por redes neuronales con propagación hacia atrás. Y por último la función de activación lineal, la cual es mayormente utilizada para la aproximación lineal, esta entrega una salida igual a la neurona de entrada que ya tiene peso asignado. Cada una de esta función de activación se explicará a detalle más adelante (Negnevitsky, 2005).

2.3. El Perceptrón

El perceptrón es un modelo lineal con relación simple de entrada-salida, el cual su función es la clasificación binaria. Además, el perceptrón utiliza la función paso de Heaviside como función de activación. La neurona basada en las neuronas biológicas fue inventada en el año 1957 por Frank Rossenblat (Patterson & Gibson, 2017).

El perceptrón de Rossenblat está basado en el modelo neuronal de McCulloch y Pitts, este modelo consiste en un combinador lineal y un limitador duro. Las sumas asignadas de las entradas se aplican al limitador, el cual produce una salida binaria, donde puede ser "+1" o "-1" esto con el objetivo de clasificar. (Negnevitsky, 2005)

El algoritmo de aprendizaje del perceptrón hace cambios de los pesos en el modelo de tal manera que todas las entradas son clasificadas de manera correcta. Este algoritmo se enfoca en separar linealmente las entradas. Este algoritmo inicializa el vector de pesos con valores pequeños o valores en 0 para inicializar el entrenamiento.

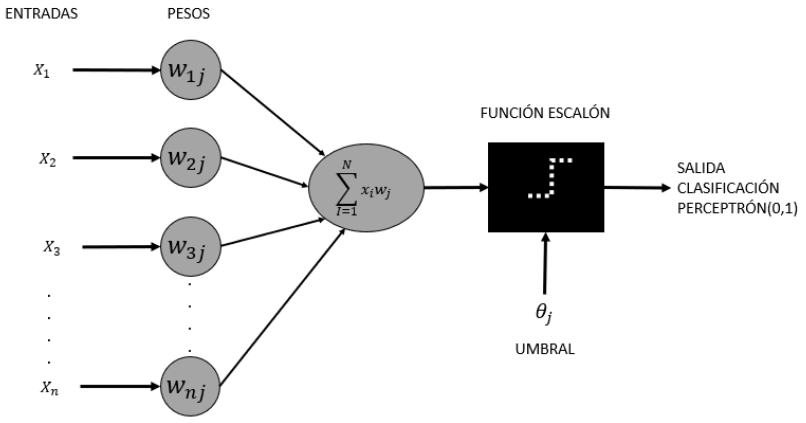

Figura 2.3: Arquitectura de Perceptrón

2.4. Redes Neuronales Multicapa

Las redes neuronales multicapa o también conocidas como el perceptrón multicapa, son redes de propagación hacia adelante, la cual se le considera multicapa porque tiene una o más capas escondidas, es decir, la red multicapa se conforma por una capa de entrada para las neuronas y al menos cuenta con una capa intermedia y al final con la capa de salida como se puede ver en 2.4. Este tipo de red se le considera propagación hacia delante por la dirección en la que viajan los datos (Negnevitsky, 2005).

El propósito de tener capas intermedias es que cada capa tenga un propósito o tarea específica, pero en conjunto las capas escondidas son las encargadas de proporcionar un patrón de salida para la red. Además, estas capas ocultas son las que detectan las características y son las que representan los valores o datos de entrada (Negnevitsky, 2005).

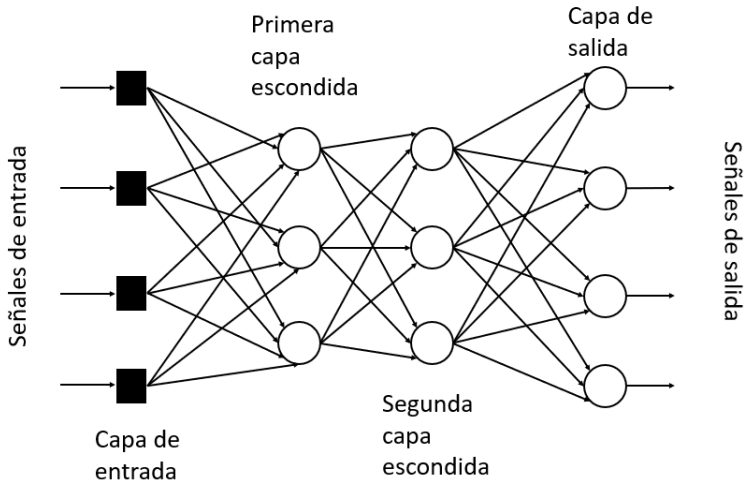

Figura 2.4: Arquitectura Red Neuronal Multicapa

Como se mencionó anteriormente, el perceptrón multicapa es similar a al perceptrón con la ventaja de que puede añadirse diferentes tipos de activación a las capas, como se muestra en la Figura 2.5.

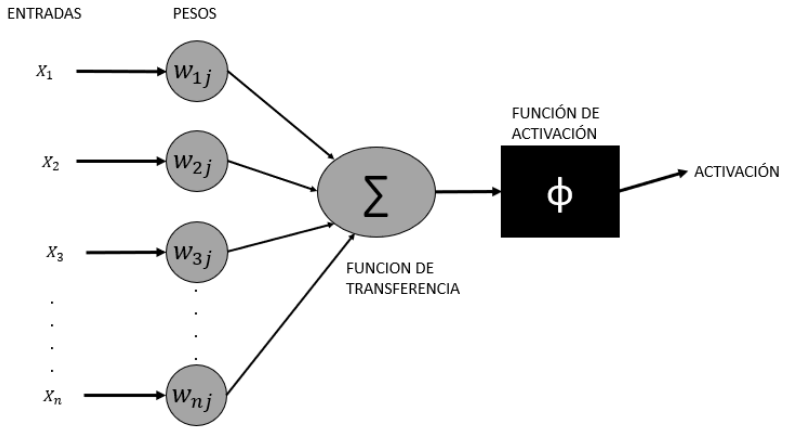

Figura 2.5: Activación de Red Neuronal Multicapa

Las redes multicapas cuentan con distintos métodos de algoritmos de aprendizaje, uno de los más utilizados es el aprendizaje por propagación hacia atrás. Este cuenta con dos fases, la primera es donde se presenta el patrón de entrenamiento a la capa de entrada, este hace que recorra capa por capa hasta la capa final o de salida. Si existe una diferencia entre el patrón con lo resultados que se espera se hace el cálculo de error y es cuando empieza la propagación hacia atrás por la red haciendo el recorrido de manera inversa y mediante esto va modificando los pesos a medida que hace la retropropagación. La Figura 2.6 se aprecia como funciona el algoritmo de aprendizaje en la red neuronal (Negnevitsky, 2005).

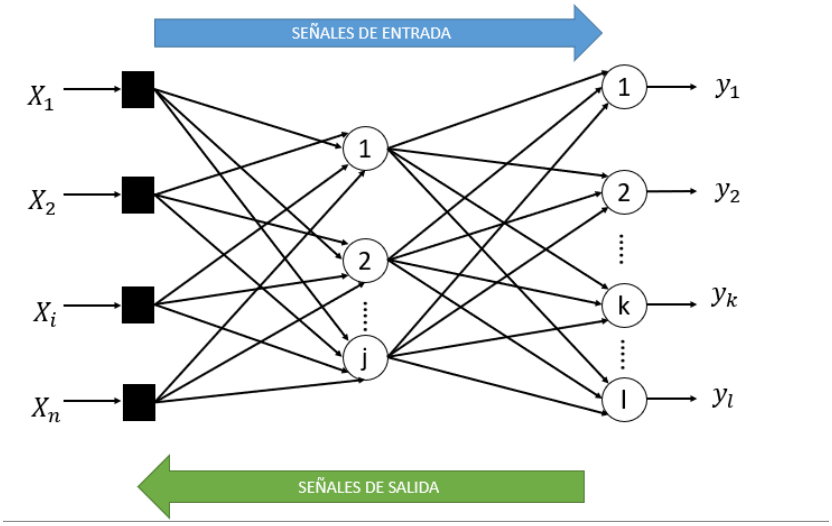

Figura 2.6: Recorrido de Retropropagación

2.5. Función de Activación

Para poder propagar la salida de los nodos de una capa a la siguiente es necesario utilizar funciones de activación, estas funciones de activación son funciones escalar-escalar que logran activar las neuronas de la red. Las funciones de activación son importantes y acompañan a las capas ocultas de la red neuronal (Patterson & Gibson, 2017).

Función Lineal

La función lineal se rige matemáticamente por la función de f(x)=Wx, la cual es la encargada de hacer la transformación lineal donde la variable independiente es directamente proporcional con la variable dependiente, de esta manera la señal pasa sin alteraciones (Patterson & Gibson, 2017).

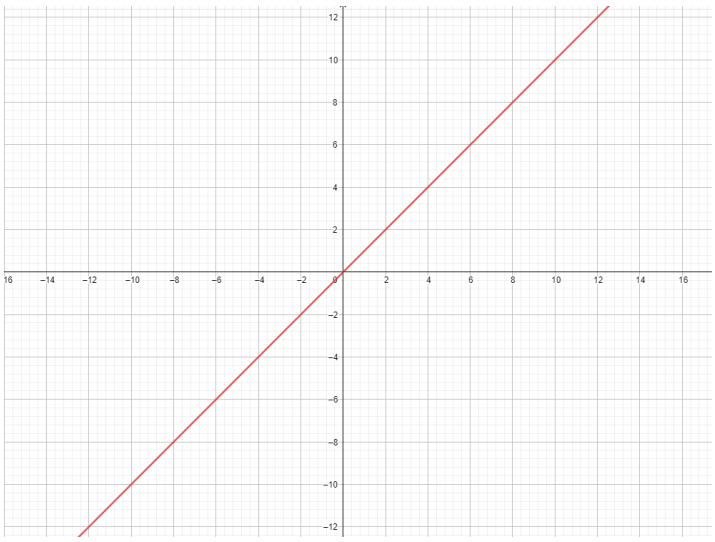

Figura 2.7: Gráfica de función Lineal

Función Sigmoide

La función sigmoide proporciona la normalización de los datos, es decir, reduce los valores convirtiéndolos a una escala entre 0 y 1, donde la mayoría de los datos se acercarán a los extremos de 0 o 1.

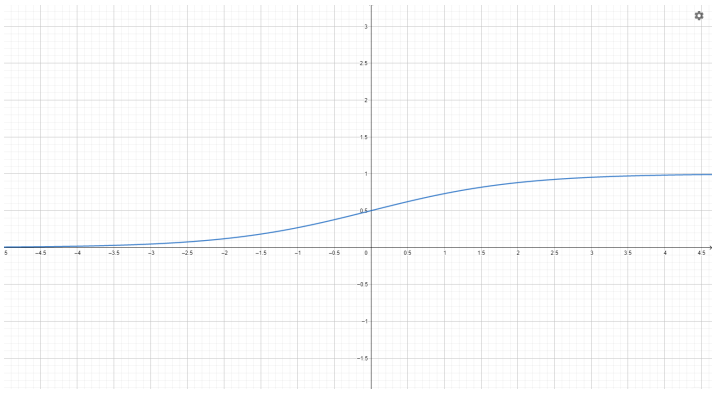

Figura 2.8: Gráfica de función Sigmoide

Función Tanh

La función Tanh o función tangente hiperbólica descrita en la Ecuación 2.2, muestra la relación que existe entre el coseno hiperbólico y el seno hiperbólico, es decir:

$$tanh(x) = \frac{sinh(x)}{cosh(x)} \tag{2.2}$$

De manera similar a la función sigmoide, la función tanh normaliza los valores entre el rango de -1 a 1.

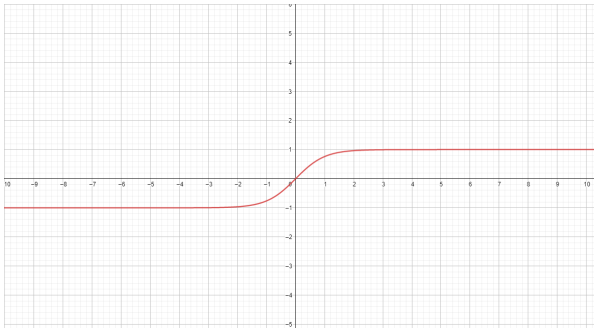

Figura 2.9: Gráfica de función Tangente Hiperbólica

Softmax

Además de la clasificación binaria, es necesario contar con la clasificación múltiple, es ahí donde la función de activación Softmax es útil. Ya que es relevante en el área de la regresión logística, debido a que se aplica a datos continuos, y usualmente esta función de activación se

encuentra en la capa de salida para clasificadores. Una de las variantes de la función Softmax que puede trabajar con miles de clasificaciones, es la Softmax jerárquica, dado que utiliza estructuras de árboles separando las etiquetas y el clasificador se entrena en cada nodo del árbol para ir agrupando cada ramificación (Patterson & Gibson, 2017).

Lineal Rectificado(ReLU)

La activación lineal rectificada descrita en la Ecuación 2.3 tiene como condición el paso de información por el nodo solo si el dato de entrada está por encima del umbral. Es decir, mientras la entrada este por debajo de 0, siempre será 0, de lo contrario la salida tendrá una relación lineal con la variable dependiente o la variable de entrada, definiéndose como:

$$f(x) = max(0, x) \tag{2.3}$$

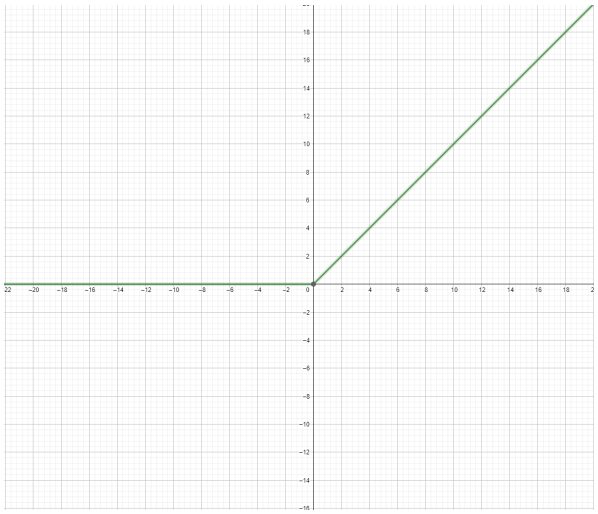

Figura 2.10: Gráfica de función Lineal Rectificado

Una de las ventajas de la función ReLU a comparación con la función sigmoidea y Tanh, es que no sufre problemas de gradiente de fuga. Además, esta función se entrena de mejor manera sin utilizar preentrenamiento (Patterson & Gibson, 2017).

2.6. Funciones de pérdida

Una de las métricas importantes para conocer que tan bien está cerca de su propósito una red neuronal es la función de pérdida. Esta función se encarga de entregar un valor basado en el error que se observa en las predicciones de las redes neuronales. Ahora bien, cada uno de estos errores obtenidos se promedian y muestra un valor único, el cual representa si la red neuronal está cerca de realizar su función adecuadamente. Se podría considerar que las funciones de pérdida son un tipo de optimizador, ya que ayuda a conocer que tan bien está aprendiendo la red neuronal al momento de realizar las predicciones mediante los errores obtenidos a través del entrenamiento (Patterson & Gibson, 2017).

Funciones de pérdida para clasificación

Existen diferentes escenarios donde las redes neuronales pueden ser utilizadas para la clasificación, como ejemplo podría ser la clasificación de datos en diferentes categorías, así como también problemas de clasificación donde el objetivo es asignar probabilidades a las clasificaciones (Patterson & Gibson, 2017).

- Pérdida de bisagra:

  Cuando se trata de redes para clasificación estrictas a extremos, es decir, tomar decisiones de 0 o 1. La función de pérdida de bisagra es la más comúnmente utilizada para la optimización de este tipo de redes. Este tipo de pérdida también suele utilizarse en modelos llamados modelos de clasificación de margen máximo (Patterson & Gibson, 2017).

  La Ecuación 2.4 describe como la pérdida de bisagra toma la decisión para la clasificación.

$$L(W, b) = \frac{1}{N} \sum_{i=1}^{N} max\left(0, 1 - y_{ij} \chi \widehat{y}_{ij}\right) \tag{2.4}$$

- pérdida logística:

La función de pérdida logística existe para aquellos casos donde las probabilidades son el punto de interés. Más allá de predecir un 0 o un 1, la idea de la función de pérdida logística es obtener la mayor tasa de probabilidad de acertar en la clase correcta para cada predicción hecha (Patterson & Gibson, 2017).

2.7. Hiperparámetros

Existen parámetros que se consideran parte del modelo y a su vez existen otros parámetros los cuales se pueden ajustar de tal manera que las redes aprendan de la mejor manera y más rápido y son denominados hiperparámetros. Estos se encargan de la optimización y la selección del modelo de entrenamiento en conjunto con el algoritmo de aprendizaje. La idea central de los hiperparámetros es mantener el modelo dentro del conjunto de datos, así evitando el subajuste y el sobre ajuste durante el entrenamiento (Patterson & Gibson, 2017).

Tasa de aprendizaje

Durante el ajuste de parámetros para la optimización y minimizar el error en la red neuronal, se necesita de un valor que se encargue de cuantificar la cantidad de ajuste que se va a estar iterando en las optimizaciones. Donde la tasa de aprendizaje se considera un coeficiente que va actualizando el vector de parámetro a medida conforme avanza el entrenamiento en conjunto a la función de pérdida. Otro parámetro que entra en juego para dar pasos grandes o cortos durante la actualización de pesos en la tasa de aprendizaje es el gradiente del error, donde la tasa de aprendizaje determina la cantidad de gradiente a utilizar para el siguiente paso del algoritmo. A medida que el error se va minimizando y el gradiente se aplana, el paso de la tasa de aprendizaje se va acortando (Patterson & Gibson, 2017).

Regularización

Como su nombre lo dice, regularización ayuda con los parámetros que están fuera de control para que no tengan un gran impacto en el entrenamiento. La idea central de este hiperparámetro es evitar el sobreajuste, el cual utiliza distintos métodos para reducir el tamaño

de los parámetros a través del tiempo (Patterson & Gibson, 2017). Conforme existen más datos de entrada como conjunto de entrada de datos, el efecto que tiene la regularización va disminuyendo, esto debido al exceso de extracción de características que se obtendrán con el conjunto de entrada de datos, de esta manera el tener mayores datos de entrada se considera un regularizador definitivo (Patterson & Gibson, 2017).

Impulso

Durante los entrenamientos existen momentos o puntos del espacio de búsqueda donde se puede atascar y jamás salir de ahí o tener un entrenamiento pobre, para esto existe el hiperparámetro de impulso. El cual se encarga de salir de esos puntos muertos en el entrenamiento, encontrando los caminos que conducen a los mínimos, de esta manera se producen modelos de mejor calidad (Patterson & Gibson, 2017).

Dispersión

Al momento de introducir datos y reconocer características, algunas no estarán presentes en los datos. La idea del hiperparámetro de dispersión es reconocer que estas características están escasas, pero son relevantes para el aprendizaje de la red neuronal. De esta manera, las neuronas son obligadas a activarse con las características escasas y así evitar que la red quede atascada (Patterson & Gibson, 2017).

2.8. Redes Profundas

Para definir que son las redes profundas, se debe hablar primero del aprendizaje profundo, dando por hecho que el aprendizaje profundo está conformado por más neuronas que una red simple donde todas estas neuronas están totalmente conectadas. Cuenta con diferentes conexiones más complejas al conectar las capas, generando así más parámetros a optimizar. Las redes profundas tienen mayor poder computacional al momento de entrenar, teniendo así problemas más complejos a resolver y por último la extracción de características en estas redes es automática (Patterson & Gibson, 2017).

Las redes profundas se conforman de 4 arquitecturas principales:

- Redes preentrenadas sin supervisión:

  Conformada por 3 arquitecturas específicas, autocodificadores, redes neuronales profundas de creencia y las redes neuronales generativas-adversarias. Las redes neuronales de creencia en su fase de preentreno utilizan capas compuestas de Máquinas restringidas de Boltzmann y redes de propagación hacia adelante para optimización. Las redes generativas-adversarias, son redes de aprendizaje sin supervisión que se entrenan de forma paralela, donde una parte es la discriminatoria, siendo una red convolucional típica que se encarga de medir similitud entre la entrada con la salida y determinar los cambios para la red generativa, donde esta red es la encargada de crear datos de entrada a partir de lo entregado de la red discriminativa. Por último, los autocodificadores son utilizados para aprender a comprimir representaciones de conjunto de datos (Patterson & Gibson, 2017).

- Redes Neuronales Recurrentes:

  Estas redes son capaces de enviar información a través de pasos de tiempo. Las redes recurrentes toman cada vector desde una secuencia de entrada de vectores y modela cada una en el tiempo, permitiendo así que la red retenga la información del vector, mientras modela otra (Patterson & Gibson, 2017).

- Redes Neuronales Recursivas:

  Las redes neuronales recursivas pueden trabajar con entradas de longitud variable. Dentro de la arquitectura que tienen las redes recursivas cuentan con una matriz de pesos compartidos y una estructura de árbol binario que permite aprender secuencias variables de palabras o partes de una imagen (Patterson & Gibson, 2017).

Redes Neuronales Convolucionales

Por último, las Redes Neuronales Convolucionales pertenecen a una de las 4 arquitecturas principales, la cual se describirá a detalle por su amplio uso en los sistemas biométricos.

Las CNN tienen como principal función aprender características mediante el uso de convoluciones. Este tipo de redes son mayormente implementados en el reconocimiento e identificación de imágenes, pudiendo así combinar la visión artificial con redes convolucionales para obtener una herramienta que se asemeje al humano al momento de identificar objetos con la vista.

Las redes neuronales convolucionales se componen a grandes rasgos de tres puntos importantes: Dato de entrada, capas de extracción de características y capas de clasificación. La Figura 2.11 se observan de manera gráfica las partes generales de una red convolucional (Patterson & Gibson, 2017).

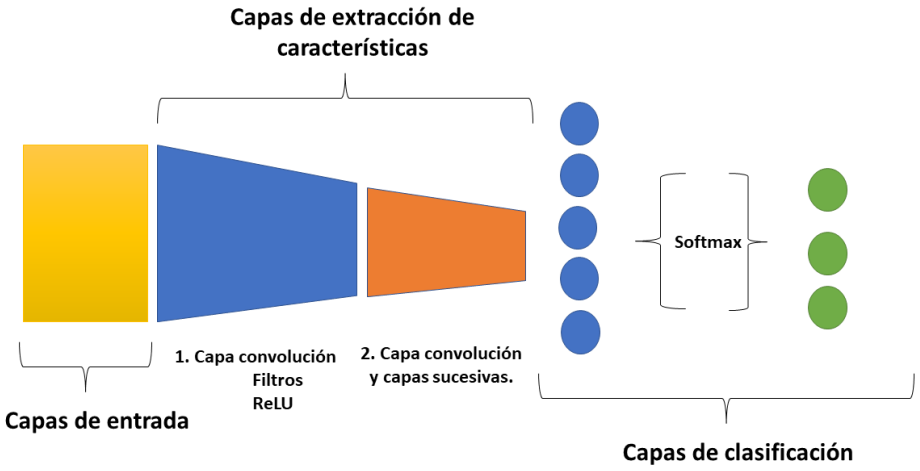

Figura 2.11: Arquitectura General Red Neuronal Convolucional

- Capa de entrada

  La capa de entrada, es donde comienza la red, recibiendo los datos en crudo. Conformado por 3 dimensiones o canales: ancho, largo y profundidad. Donde el ancho y largo que tienen como atributos son los píxeles y la profundidad se rige por los canales RGB.

- Capas de extracción de características

  Conformado principalmente por capas convolucionales y capas “pooling”, encargadas de encontrar las características en las imágenes y construir a partir de ellas la entrada para la siguiente capa.

- Capas de extracción de características

Por parte de la capa de clasificación tenemos una o más capas totalmente conectadas a las capas anteriores. Aquí se producen las salidas de matrices de dos dimensiones donde se tiene los valores de la imagen con su clasificación o predicción hecha por la red.

Capas de extracción de características

- Capas de Convolución:

  Las capas de convolución se consideran la parte más importante de la arquitectura de la red neuronal. Durante el proceso de convolución se realiza un proceso matemático donde se recorre los datos de entrada con un kernel, reduciendo el tamaño de los datos por región utilizando el producto punto en conjunto con los pesos de activación entre cada capa. En la Figura 2.12 se observa como se hace el proceso de convolución.

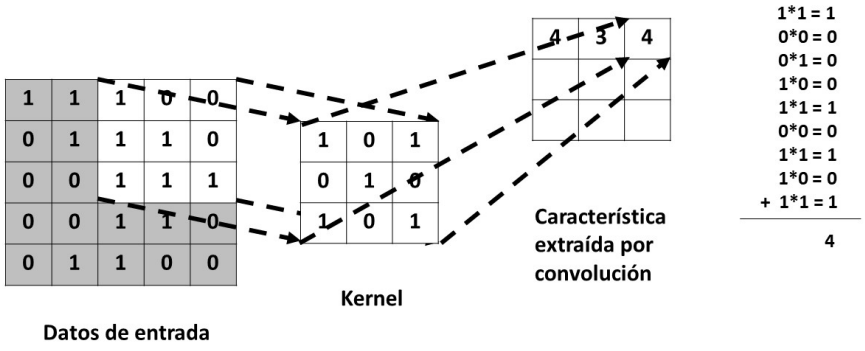

Figura 2.12: Operación de convolución

Existen parámetros e hiperparámetros que caracterizan a las capas convolucionales, esto con la finalidad de mejorar el entrenamiento. Los parámetros que más resaltan son: los filtros, en este caso el kernel que se aprecia en la Figura 2.12. Por otro lado, los mapas de activación, los cuales son la representación visual de los números de activaciones en varias capas de la red, los parámetros compartidos son utilizados para el controlar en número total de parámetros (Patterson & Gibson, 2017).

Por último, se tiene la capa de hiperparámetros para la capa convolucional el cual se conforma por, el tamaño de filtro, profundidad de salida, paso, y Zero-Padding.

- Tamaño de filtro:

  Como su nombre lo indica, el filtro o kernel a utilizar necesita un tamaño para realizar el proceso de convolución, para esto se le da un tamaño al filtro para empezar a realizar el recorrido entre todos los datos de entrada.

- Profundidad de salida:

  Es el encargado de controlar el número de neuronas de la capa convolucional, la cual es conectada a la misma área del volumen de entrada.

- Paso:

  Este hiperparámetro es el que determina cuantos valores va a recorrer el filtro para aplicar la convolución. Mediante este hiperparámetro se van generando nuevas columnas de salida de profundidad, dependiendo así del número de pasos que, de dentro de la entrada de datos, será el número de columnas de salida.

- Zero-Padding

  El Zero-Padding tiene la función de controlar el tamaño de espacios que se va a encontrar en la salida de la convolución.

- Capas Pooling

  Ahora bien, para evitar el sobreajuste u *overfitting*, es necesario ir reduciendo progresivamente el tamaño de entrada, de esta manera después de una capa convolucional se va añadiendo capas pooling. Estas capas actúan de manera independiente en cada reducción de profundidad para los datos de entrada (Patterson & Gibson, 2017).

  Para realizar el reescalado o redimensionar los datos de entrada, se utiliza el Max-pooling, utilizando así comúnmente un tamaño de filtro de 2x2. La operación de Max pooling se encarga de tomar el número más grande dentro del filtro sin afectar la profundidad de los datos.

- Capa Totalmente conectada (Fully Connected)

Es la capa encargada de calcular los valores de las clases que se utilizan como salida de la red neuronal. La salida es un vector del tamaño de 1x1xN donde N es el número de clases a evaluadas.

VGGVox

La Red Convolucional VGGVox del trabajo (Nagrani et al., 2017) es una red diseñada para la identificación de voces, de esta manera la red fue estructurada para clasificación de multiclase, donde cada persona corresponde a una. La arquitectura está basada en la VGG-M, una red convolucional con buen rendimiento para la clasificación de imágenes, de esta manera se adaptó para entradas de espectrogramas, una característica extraíble de los audios. Las modificaciones a la red fueron en una de las capas llamadas *Fully-Connected* 6 la cual se reemplazó por dos capas, una de *Fully Connected* de 9x1 para el dominio de las frecuencias y una capa *average pooling* de 1xn, donde n varía dependiendo del tamaño del audio en segundos, logrando así que la red soporte diferentes tamaños de audio.

Las modificaciones realizadas lograron reducir los parámetros de entrenamiento de 319M de la VGG-M a 67M en la red VGGVox.

2.9. Biometría

Existen rasgos únicos que distinguen a cada individuo uno de otro, clasificadas en categorías fisiológicas y de comportamiento denominado rasgos biométricos. La biometría se puede definir como un método para medir las características únicas de los individuos. Actualmente, estos rasgos pueden ser utilizados para la autentificación e identificación de personas (Arya & Bhadoria, 2019).

En este apartado describiremos los tipos de rasgos biométricos, que es un sistema de autentificación biométrica, un esquema general de como es el proceso de identificación de estos sistemas, se describirá a detalle sobre el reconocimiento de voz y los tipos de características que existen en la voz.

Tipos de rasgos biométricos

Como se mencionó anteriormente, los rasgos biométricos se pueden clasificar en dos tipos: fisiológicas y de comportamiento (Gayathri et al., 2019).

Características físicas:

A continuación se describirán los rasgos físicos más utilizados actualmente y una descripción de ellos.

- Huella dactilar:

  Considerado uno de los rasgos más explorados y utilizados actualmente como método de autentificación gracias a su bajo costo y alto desempeño. Su proceso es simple, se toma una captura, es decir, una imagen de la huella de una persona, se realizan preprocesamiento de mejora de imagen y posteriormente se hace la extracción de características. Al final se guarda el vector de características extraído de la huella (Joshi et al., 2018).

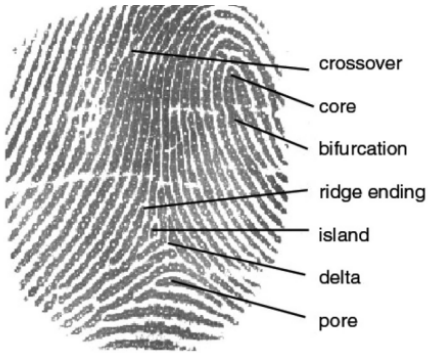

Figura 2.13: Estructura de la huella dactilar(Marasco & Ross, 2014).

- Iris:

  Los sistemas de reconocimiento de iris son los más exactos dentro de la seguridad biometrica. Este proceso puede utilizar distintos extractores de características como el "*Eyelid*" y "*eyeslashes*" mediante el uso de iluminación infrarroja (Ammour et al., 2018).

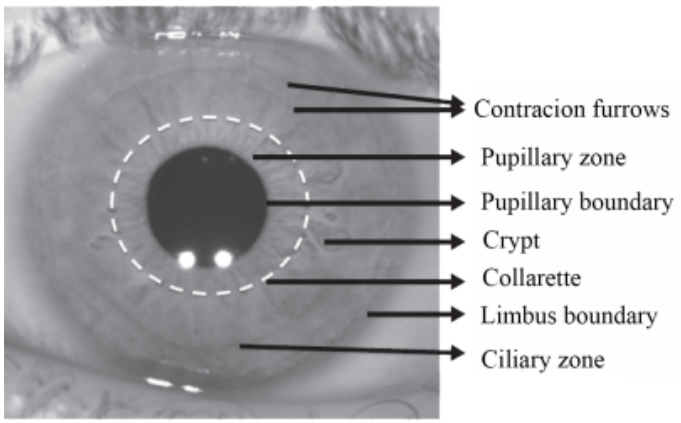

Figura 2.14: Estructura del iris(Daouk et al., 2002).

- Rostro:

  Uno de los sistemas más importantes para el área de seguridad pública. Para la extracción de rostro es necesario contar con una cámara de buena calidad para tener mayor precisión a la hora de realizar la clasificación. Mediante la detección de ojos, nariz y boca, se realiza el proceso de autentificación (Zulfiqar et al., 2019).

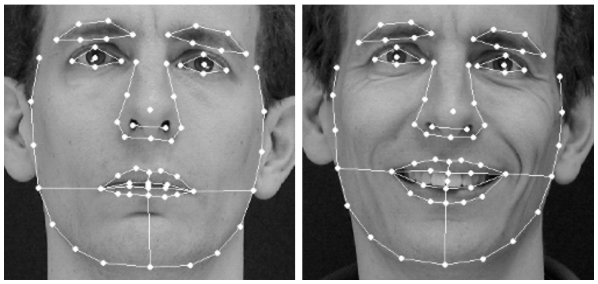

Figura 2.15: Reconocimiento de rostro(Jain & Li, 2011).

- Palma de la mano:

  Mediante el uso de un scanner se captura las características de la palma, utilizando un sensor térmico u óptico. Las crestas y arrugas son extraídas con métodos de línea, subespacio y estadístico para posteriormente ser procesados y extraer el vector de características (Kong et al., 2009).

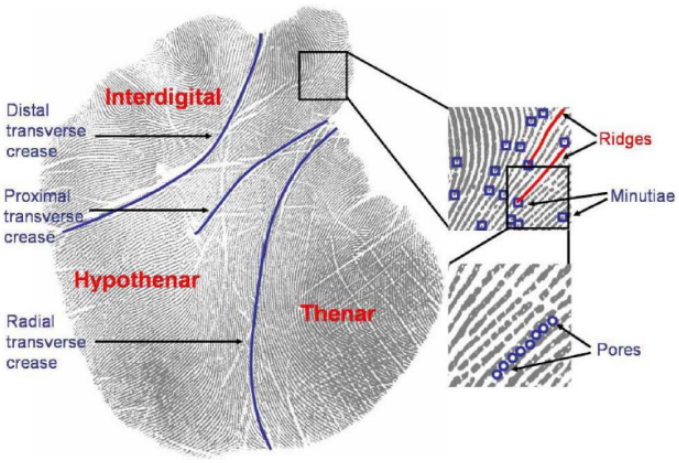

Figura 2.16: Estructura de palma de la mano(Dai & Zhou, 2010).

- ADN:

  Dentro del código del ADN existen 4 características importantes: adenina, guanina, citosina y timina. Este tipo de autentificación se utiliza mayormente en el área forense y médica. Mediante el uso de la saliva, sangre o cabellos y mediante un laboratorio se hace la extracción del ADN (Hashiyada, 2011).

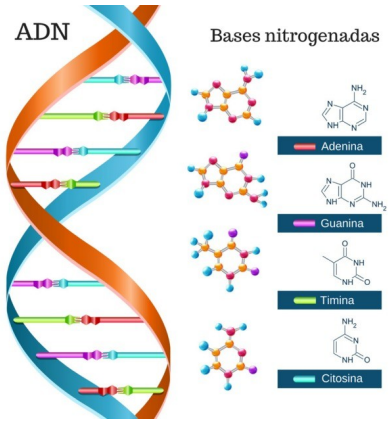

Figura 2.17: Estructura del ADN

En la Tabla 2.1 se muestran los rasgos descritos anteriormente y adicionalmente se agregan parámetros medibles y áreas de aplicación.

Tabla 2.1: Tabla de rasgos biométricos fisiológicos y parámetros medibles

Rasgo Biometrico	Parametros medibles	Aplicaciones del rasgo
Huella dactilar	patron de crestas, puntos de minucias, poros	Sistemas de accesos de seguridad, entrada y salida para trabajadores
Iris	separación de parpados y petañas	Seguridad bancaria
Rostro	Extracción de ojos, nariz y boca	Sistemas de autentificación y vigilancia
Huella de palma	Lineas principales, arrugas y crestas	Areas de criminalistica
ADN	Secuencia de codón, aminoacidos y proteinas	Casos forenses, genetica

Características de comportamiento:

En este tipo de rasgos se desea encontrar el patrón en los aspectos humanos de comportamiento.

- Andar humano:

  Gracias a la visión computacional, uno de los rasgos biométricos que han tomado gran importancia es el de reconocimiento del andar humano, ya que mediante la forma de caminar y el ciclado de movimiento se encuentra el patrón mediante técnicas de aprendizaje máquina. Aplicado mayormente en sistemas de videovigilancia para temas de seguridad (Zhang et al., 2019).

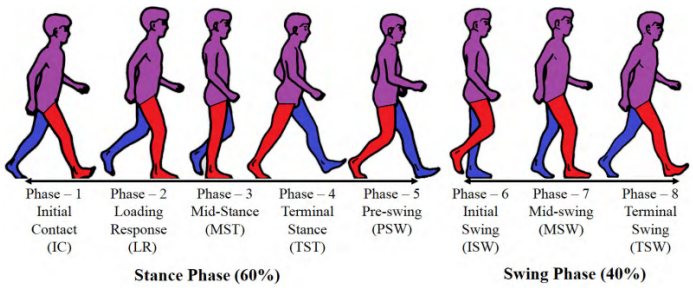

Figura 2.18: Ciclo del caminado humano(Singh et al., 2018).

- Voz:

  Mediante la toma de muestra de la voz y la utilización de distintos extractores de características como los Coeficientes Cepstrales de Mel, Espectrogramas, Filtro de Bancos, entre otros extractores que ayudan encontrar patrones de reconocimiento mediante la señal del audio utilizado en el dominio de las frecuencias (Nilu et al., 2018).

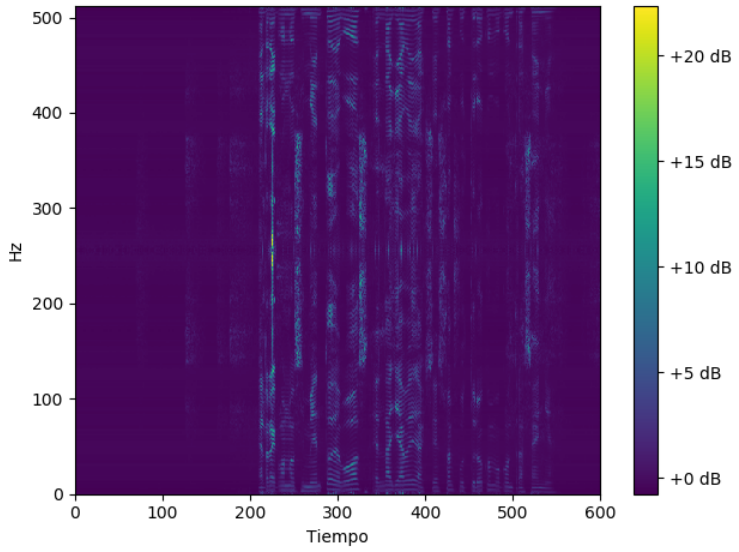

Figura 2.19: Visualización de la voz en el dominio de la frecuencia

- Escritura:

  Cada persona tiene una distinta forma de escribir, mediante el uso de tabletas digitales es posible extraer distintos parámetros como contar segmentos, número de eventos de pluma arriba y pluma abajo, duración al escribir, presión, separación de letra, etc. (Zhu et al., 2000).

- Firma:

  De la misma manera que la escritura, la firma es capturada mediante el uso de tableta, midiendo así la presión coordenada de cada punto de escritura. Este tipo de rasgo es mayormente implementado en sistemas de compras electrónicas y oficinas de patentes (Querini et al., 2015).

De la misma manera que en los rasgos fisiológicos, en la Tabla 2.2 se muestra la información de los rasgos de comportamiento de manera concentrada.

Tabla 2.2: Tabla de rasgos biométricos de comportamiento y parámetros medibles

Rasgo Biometrico	Parametros medibles	Aplicaciones del rasgo
Andar humano	altura, largo de paso y velocidad	area medica, vigilancia
Voz	contorno de energia, frecuencia, valores de meseta de energia	autentificación en centro telefonicos
Escritura	separación de escritura, presion, inclinación	aplicaciones forenses
Firma	velocidad la firma, angulo entre trazos	Oficinas de patentes y compras electronicas
ADN	Secuencia de codón, aminoacidos y proteinas	Casos forenses, genetica

2.9.1. Sistema Biométrico

Retomando los rasgos biométricos, al momento de ser implementados como métodos de autentificación, es necesario desarrollar un sistema que, tenga un sensor apropiado para el rasgo a utilizar para capturar el dato de entrada y posteriormente realice sus respectivos preprocesamientos y sea capaz de extraer las mejores características representativas del rasgo de entrada, generando así un "*embedding*" o vector de características.

Un sistema biométrico se puede describir de manera general mediante 4 etapas. Una vez se cuente con el sensor necesario, la primera etapa consta del registro del usuario o ("*Enrollment*") almacenando así una muestra biométrica. En la segunda etapa se extraen las características de la muestra, se obtiene el parámetro a medir del dato de entrada mediante un algoritmo llamado Extractor de características o ("*Feature Extractor*") y se guarda en una base de datos como una plantilla con un nombre identificador.

Para la tercera etapa, entra la autentificación de usuario, donde se presenta otra muestra en el sensor. Esta muestra no es almacenada sino, inicia un proceso denominado "*query*" en el cual el sistema compara el dato de entrada con el registrado como plantilla. Llegando así a la cuarta etapa donde se realiza un ("*matcher*") o comparador, el cual retorna un valor representativo para medir la similitud entre la muestra en la base de datos y la muestra de entrada. Una vez obtenido el valor de similitud, se decide si el sistema acepta la autentificación, siempre y cuando este dentro del umbral de aceptación (Jain & Nandakumar, 2012).

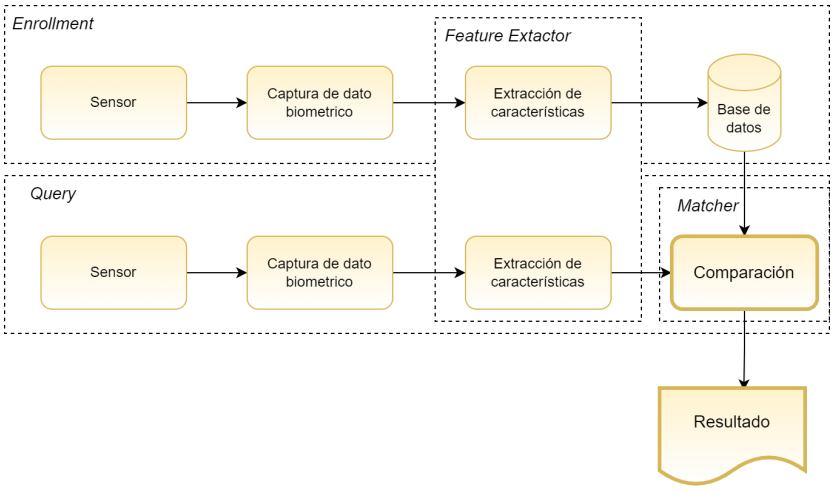

Figura 2.20: Diagrama general de un sistema de autentificación biométrico

Criterios de evaluación de calidad para un sistema biométrico

Dentro de los sistemas biométricos existen diversos aspectos de calidad que se pueden evaluar en ellos, en el trabajo de (Rui & Yan, 2018) mediante una exhaustiva investigación entre diversas fuentes, proponen 5 aspectos importantes a evaluar dentro de los sistemas biométricos, los cuales son:

- Exactitud:

  Representado por 3 parámetros dentro de estos sistemas, FAR (*False Acceptance Rate*) es decir el porcentaje de falsos positivos, FRR (*False Rejection Rate*) o también el porcentaje de falsos negativos y EER (*Equal Error Rate*) el cual representa el promedio de los parámetros descritos anteriormente.

- Eficiencia:

  Esta característica indica el tiempo que toma el sistema en autentificar a una persona mediante el rasgo biométrico utilizado.

- Usabilidad:

  El criterio de usabilidad se evalúa con distintos factores, universabilidad, unicidad, permanencia y aceptabilidad.

- Seguridad:

  Existen métodos de proteger de ataques al rasgo biométrico utilizado, así evitar reproducción o falsificación de este.

- Privacidad:

  El criterio de privacidad está enfocado en que tan protegido está el rasgo biométrico para ser obtenido en la vida real. Dentro de la privacidad se toma en cuenta una característica a evaluar el MSR (*Mission Success Rate*) midiendo la posibilidad de resistir ataques y proteger la privacidad del dato biométrico.

Mediante estos 5 criterios, se desarrolla la Tabla 2.3 en la cual se dan 3 valores para clasificar el sistema: Alto, Medio, Bajo y los criterios a considerar para entrar en cada una de estas categorías.

Tabla 2.3: Evaluación de calidad en sistemas biométricos

Criterio	Niveles		
	Alto	Medio	Bajo
FAR, FFR o EER	La evaluación de FAR, FRR o EER son menores al 3 %	La evaluación de FAR, FRR o EER estan entre el 3 % y el 10 %.	La evaluación de FAR, FRR o EER son mayores al 10 %.
Eficiencia	Duración de identificación menos de 1 segundo. O se menciona que el algoritmo tiene tiempo bajo en costo computacional, siendo capaz de implementarse en dispositivos moviles.	Duración de identificación de 1 a 3 segundos. O el metodo necesita un proceso de entrenamiento y aprendizaje, pero el costo computacional es medio, el cual es posible a implementarlo en dispositivo movil pero no muy adecuadamente.	Duración de identificación mayor a 3 segundos. O se menciona que el algoritmo tiene tiempo alto en costo computacional, por lo cual no es adecuado implementarse en un dispositivo movil.
Universalidad	Todo mundo cuenta con el rasgo biometrico, y no es afectado por discapacidades, enfermedades y accidentes.	Existe una pequeña probabilidad que el rasgo biometrico puede ser afectado por algun accidente o enfermedad.	Una gran cantidad de usuarios no cuentan con el rasgo biometrico.
Unicidad	Todos los humanos tienen diferente comportamiento en el rasgo biometrico.	La caracteristica biometrica es diferente en una gran escala.	La caracteristica biometrica es solo diferente en una pequeña escala.
Permanencia	El rasgo biometrico no cambia a lo largo de la vida del usuario	El rasgo biometrico no cambian drasticamente en muchos años	El rasgo biometrico puede cambiar significativamente en un periodo corto
Aceptabilidad	El rasgo biometrico a sido ampliamente utilizado en sistemas de autentificación en la industria y negocios.	La autentificación biometrica a sido implementada, pero no ha sido ampliamente usada.	Hay pocos ejemplos de aplicaciones practicas.
Seguridad	Solución en la seguridad propuesta con pruebas demostradas	La caracteristica biometrica tiene una particularidad relativamente dificil de atacar. Se ha discutido poco el tema de seguridad.	Pocos estudios en el tema de seguridad. La caracteristica biometrica no es segura.
MSR	Porcentaje de MSR mayor o igual a 90 %	Porcentaje de MSR mayor o igual a 50 % pero menor que 90 %	Porcentaje de MSR menor que 50 %

Para calcular el desempeño de los sistemas biométricos, como la exactitud, la cual tiene la finalidad de medir la aceptación/rechazo de una persona, se utilizan 3 métricas distintas, FAR, FRR o EER (Petrovska-Delacrétaz et al., 2009). Las métricas están definidas cada una como:

- *False Acceptance Rate* (FAR): Es la métrica que determina la probabilidad de que una persona ajena sea detectada como dentro del sistema (de Martin-Roche et al., 2001).

  Para determinar el valor de FAR, se utiliza la Ecuación 2.5

$$FAR = \frac{Numero\,de\,personas\;falsas\;aceptadas}{Numero\,total\,de\,personas\;Falsas} x100 \tag{2.5}$$

- *False Rejection Rate* (FRR): Es la encargada de medir la probabilidad de un usuario registrado ser rechazado por el sistema, como si no perteneciera a él (de Martin-Roche et al., 2001).

El FRR se calcula con la Ecuación 2.6 mostrada a continuación:

$$FRR = \frac{Numero\ de\ personas\ verdaderas\ rechazadas}{Numero\ total\ de\ personas\ verdaderas} x100 \tag{2.6}$$

- *Equal Error Rate* (EER): El valor determinado donde los porcentajes de FAR y FRR son igual (de Martin-Roche et al., 2001).

2.9.2. Reconocimiento de voz

La principal tarea del reconocimiento de voz es autentificar la identidad de una persona mediante el uso de las distintas características de la voz. El reconocimiento de voz engloba diferentes características de comportamiento como las emociones y, por otro lado, las fisiológicas, es decir, el acento, el tono, intensidad, todo depende de tu estructura interna física (Minaee et al., 2019).

El reconocimiento de voz se puede dividir en dos tipos:

- Identificación de voz

  La identificación de voz, se tiene una voz desconocida como dato de entrada y esta es comparada con una base de datos de personas preestablecidas, es decir, se realiza la comparación de 1 contra N personas y se identifica a la voz que más se adapte dentro de la base de datos. En la Figura 2.21 se puede ver gráficamente como es el proceso de identificación (Tandel et al., 2020).

Figura 2.21: Identificación de voz

- Verificación de voz

A diferencia de la identificación de voz, los sistemas de verificación autentifican mediante un usuario y el dato de entrada, es decir, se conoce la persona que quiere entrar, se introduce la voz y sé válida si realmente es la persona que desea ingresar, haciendo una comparación de 1 a 1 (Tandel et al., 2020), como se puede ver en la Figura 2.22.

Figura 2.22: Verificación de voz

2.9.3. Extractores de características en la voz

Existen diversos extractores de características para una señal de audio, el cual es esencial para poder identificar los patrones al momento de querer realizar reconocimiento de voz. A continuación se analizarán tres de los principales

- Espectrograma

  La energía acústica se puede representar visualmente mediante las frecuencias a través del tiempo, a esto se le denomina espectrograma, donde el eje horizontal representa el tiempo, vertical las frecuencias y la potencia es representada como la intensidad en cada punto de intersección de tiempo-frecuencia. Las frecuencias son obtenidas mediante el uso de las transformadas de Fourier, el cual pasa el audio del dominio del tiempo al dominio de las frecuencias.

  Para realizar la extracción del espectrograma es necesario realizar ciertos preprocesamientos que ayuden a mejorar el audio para una mejor extracción de características. Es esencial contar con 2 tipos de preoprocesamientos para esto, la Fragmentación de audio (*Audio Framming*) y el deslizamiento de ventana (*Windowing*)

  - Fragmentación de audio (*Audio Framming*):

Dentro de una señal de audio existen variaciones, es decir, no es estacionaria a lo largo del tiempo, pero si se divide estos audios en pequeñas particiones sabemos que el audio se vuelve estacionario. Este proceso es ideal para poder realizar la transformada de Fourier en pequeños intervalos de la señal obtenida del audio.

Para este proceso se debe contar con dos variables, una el tamaño de ventana y otra el deslizamiento D, para poder realizar solar el audio con el anterior y el siguiente.

- Deslizamiento de ventana (*Windowing*):

  Para poder realizar el proceso de calcular la Transformada de Fourier, es necesario que se cumpla una condición en la cual el audio sea cíclico, en otras palabras, se espera que existan periodos completos en el audio. Al momento de realizar el Framming este principio de tener un ciclo, no se cumple, es por eso que es necesario filtrar la señal para mitigar o atenuar el inicio y final del cada cuadro extraído del audio para así posteriormente realizar la Transformada de Fourier. Al proceso de atenuación se le conoce como Windowing, el cual se basa en multiplicar cada cuadro o frame por una ventana, la cual es conocida como función Hamming. En la Figura 2.23 se puede observar la forma de la función Hamming.

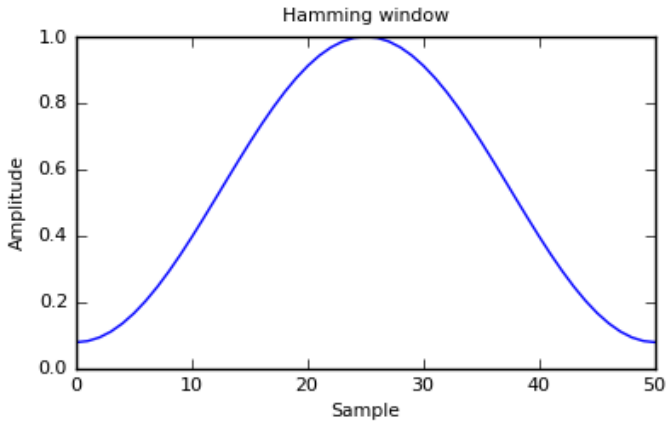

Figura 2.23: Ventana Hamming

Ahora bien, esta función está descrita por la Ecuación 2.7:

$$w\left(n\right) = 0.54 + 0.46cos\left(\frac{2\pi n}{M-1}\right) \tag{2.7}$$

- Coeficientes Cepstrales de Frecuencias de Mel

  Los Coeficientes Cepstrales de Frecuencias de Mel o MFCC por sus siglas en inglés, están enfocadas en la percepción auditiva humana, es decir, toma como máximo las frecuencias pico que logra captar el oído humano, la cual es de 1000Hz. Los MFCC cuenta con dos tipos de filtrado, el primero está espaciado linealmente en las frecuencias bajo de 1000Hz y el segundo se encuentra en frecuencias espaciadas logarítmicamente por encima. La importancia de los Coeficientes Cepstrales de Frecuencias de Mel es capturar las características importantes de la fonética de la voz humana (Bala et al., 2010). La ecuación matemática que define a los MFCC está en la Ecuación 2.8

$$Mel\left(f\right) = 2595l \cdot log\left(1 + \frac{f}{700}\right) \tag{2.8}$$

Gráficamente la relación que existe entre la escala de Mel y el dominio de las Frecuencias se puede observar en la Figura 2.24

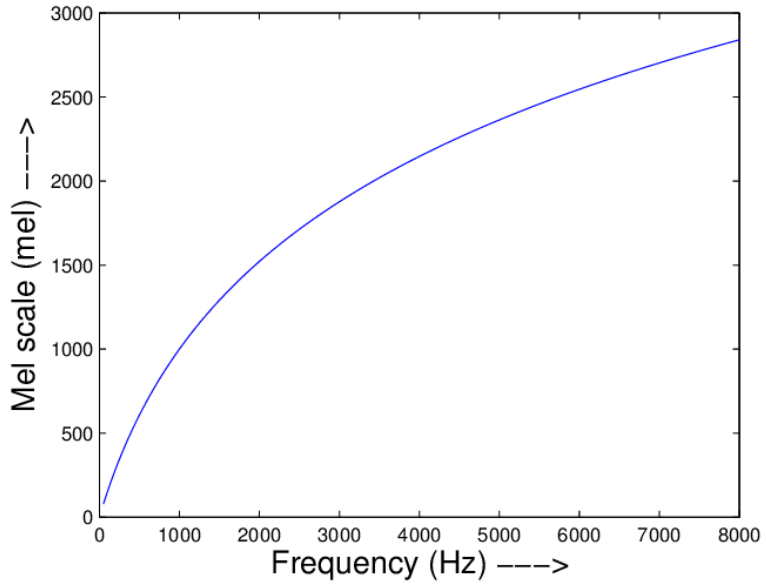

Figura 2.24: Relación Escala de Mel vs Frecuencias

2.10. Tecnologías

En la actualidad existe distintas herramientas que facilitan el desarrollo de sistemas en aprendizaje-máquina, redes neuronales y redes profundas. Además de contar con distintos lenguajes de programación de alto nivel, para principiantes y expertos. Es por eso que en este apartado se darán a conocer las herramientas que se utilizan comúnmente en la investigación que utilizan inteligencia artificial(IA).

2.10.1. Python

Una de los de lenguaje de programación principales en las que se desarrollan sistemas de IA actualmente es Python, el cual fue diseñado para ser interactivo y orientado a objetos. Este cuenta con módulos, excepciones, tipado dinámico, convirtiendo este lenguaje a uno de alto nivel, siendo un lenguaje de muchos paradigmas, como orientado a objetos, procedimental y funcional (van Rossum, 1991).

Python se caracteriza por contar con una sintaxis clara, y utilizarse como un lenguaje de extensión para aplicaciones que necesitan interfaz programable. Por último, se considera un lenguaje portátil, ya que funciona en todos los sistemas operativos. Actualmente, Python se encuentra en la versión 3 es de código abierto y es muy popular, haciendo que cuente con múltiples "frameworks" y bibliotecas que cuentan con muchos módulos para desarrollar muchas aplicaciones, en especial en el área de la inteligencia artificial (van Rossum, 1991).

2.10.2. Keras

Es una biblioteca de aprendizaje profundo basado en el lenguaje de Python, el cual va de la mano con TensorFlow. La idea central de Keras es poder utilizar sus herramientas para el desarrollo rápido de los experimentos, ya que cuenta con códigos dentro de su API que facilita la implementación del aprendizaje profundo, ayudando al área de investigación de Inteligencia artificial (Chollet, 2015).

Las ventajas de Keras es que ayuda al desarrollador enfocarse en los problemas que realmente importa, es potente, ya que tiene buen rendimiento y una escalabilidad a nivel indus-

trial, siendo utilizado en la NASA, youtube, Waymo, etc (Chollet, 2015).

2.10.3. TensorFlow

TensorFlow es una herramienta diseñada tanto para expertos como para principiantes, el cual facilita la implementación de modelos de aprendizaje automático. Esta herramienta ofrece distintos niveles de abstracción para elegir el que más se adapte a las necesidades del proyecto en conjunto con la API de Keras que se mencionó anteriormente(Martín Abadi et al., 2015).

Tiene la ventaja de poder trabajar con distintos lenguajes de programación, como Python, java, javascript, C++ etc. La biblioteca de TensorFlow fue desarrollada por Google para construir y entrenar redes neuronales y así reconocer patrones y correlacionesMartín Abadi et al., 2015.

2.10.4. PyTorch

La biblioteca de PyTorch está diseñada para el desarrollo de aprendizaje profundo mediante el uso de GPUs y CPUs, cuenta con similitudes a TensorFlow y Keras, esta biblioteca está basada en Torch la cual está diseñada para el desarrollo en el lenguaje C (Paszke, s.f.).

Su principal enfoque es el desarrollo de visión artificial y procesamiento de lenguajes naturales. Las principales características que cuenta PyTorch son la computación de tensores, la cual puede hacer uso de GPUs para la aceleración de los proyectos y está enfocada en redes neuronales profundas (Paszke, s.f.).

Capítulo 3

Estado del Arte

Durante los años 30, Francés McGehee se inspiró en un caso de secuestro y asesinato para el desarrollo del reconocimiento de voz, ya que uno de los afectados reconoció la voz del secuestrador. La idea de McGehee fue realizar una investigación de que tan fiable es el oído humano, que posteriormente esto daría partida a un tipo de investigación forense y psicológica (Hanifa et al., 2021).

Actualmente, la investigación de reconocimiento de voz sigue gracias a la inteligencia artificial, enfocado en técnicas de aprendizaje-máquina y aprendizaje profundo. Esto debido a que se han explorado distintos extractores de características tales como la Transformada Discreta de Wavelet donde principalmente sé utilizando los Modelos Gaussianos Mixtos (GMM por sus siglas en ingles) y el Perceptrón Multicapa (MLP por sus siglas en ingles) para el reconocimiento de voz (Hanifa et al., 2021). Por otro lado, tenemos los Coeficientes Cepstrales en las Frecuencias de Mel utilizándose en los mismos modelos anteriores y además en DNN, SVM y CNN (Hanifa et al., 2021). Existen otras técnicas de extracción con Vectores X, espectrogramas, características espectrales dinámicas normalizadas, Coeficientes cepstrales temporales basados en la energía de Teager, además de las combinaciones de estas (Hanifa et al., 2021).

El reconocimiento de voz puede estar clasificado en dos grupos, la identificación y verificación, donde la identificación se encarga de procesar una voz desconocida y compararla entre voces ya establecidas. La voz es verificada como a la que mejor se adapte. De tal modo que se ingresa una voz desconocida al sistema y este se va a encargar de obtener el nombre o de identificar quien es el usuario.

Ahora bien, la verificación de voz, al igual que la identificación, la entrada, es una voz desconocida, pero con la condición de tener otro parámetro de entrada que es el nombre de usuario con el que desea verificarse. La voz de entrada se hace comparativa con la voz ya previamente registrada ligada al usuario y como salida se obtiene un resultado de sí o no, es el usuario que desea ingresar.

Los sistemas de reconocimiento de voz se pueden visualizar o separar en dos distintos enfoques. Por un lado, está el enfoque de los modelos tradicionales, llamados así, ya que pertenecen a modelos matemáticos-estadísticos, antes de la llegada de las redes neuronales profundas. Estos modelos son los HMM (*Hidden Markov Model*), GMM (*Gaussian Mixture Model*) algoritmos basados en SVD (*Singular Value Decomposition*), DCT (*Discrete Cosine Transform*), Enfoque de agrupamiento iterativo, índice de probabilidad, entre otros derivados de los mencionados anteriormente. Ahora, por parte del enfoque del Aprendizaje profundo, algunos ejemplos son, DNN (*Deep Neural Network*), CNN (*Convolutional Neural Network*), SVM (*Support Vector Machines*) y entre otros derivados de estos mismos.

En los siguientes apartados se tratarán diversos artículos dentro de la literatura del estado del arte para el reconocimiento de voz, en el cual se separaron por tipo de técnica de aprendizaje máquina.

3.1. Trabajos que utilizan Métodos Estadísticos

En el trabajo (Dhakal et al., 2019) proponen la utilización de la máquina de soporte de vectores, el cual consiste en tratar el reconocimiento de voz como un problema de clasificación binaria. Durante el proceso de reconocimiento de voz se aplica el clasificador entrenado en distintos puntos para reconocer si la voz coincide o no. El modelo utiliza la función de base radial (RBF por sus siglas en ingles), ya que mapea de forma no línea las muestras en un espacio de mayor dimensión. Por otra parte, tiene menores dificultades numéricas y tiene menos hipeparámetros que el kernel polinómico.

Existen algunas técnicas más recientes donde se proponen variantes y mejoras de extractores de características, ya que estas definen que tan bien van a detectar a la persona que habla. En el artículo (Jahangir et al., 2020) se menciona como se combina el uso de los MFCC con

características basadas en el tiempo, convirtiéndola en MFCCT, con la finalidad de mejorar la precisión de los sistemas de Identificación de Voz.

Existen modelos basados en otras técnicas de aprendizaje máquina fuera de las redes profundas, un ejemplo es en el trabajo (Das & Nahar, 2016), donde se utiliza una técnica de Modelo Oculto de Markov. Este modela el algoritmo como un doble proceso estocástico en el que los datos en observación son el resultado de pasar un proceso oculto por un segundo proceso. Es decir, ambos procesos se caracterizan solo de un proceso que se observa. En este modelo para la extracción de características se utiliza la técnica de Coeficientes Cepstrales de Frecuencia de Mel y mediante la utilización de Cuantificación Vectorial se realiza el entrenamiento y clasificación de las características.

Por parte de sistemas implementados en un área en específico, los autores en el artículo (Tamoto & Itou, 2019) donde el estudio está centrado en la verificación de la persona mediante la voz dentro del ambiente de un vehículo o en la conducción. Uno de los principales problemas durante el estudio, es el ruido ambiental dentro un auto en movimiento. Para mitigar este problema, se utiliza la sustracción espectral y un filtrado de baja frecuencia reduciendo así el ruido. Mediante estas técnicas redujeron alrededor de un 66 % en la tasa de falsos rechazos. Para el sistema se utilizó un Modelo Gaussiano Mixto Posteriograma para la variabilidad entre hablantes.

Dentro de la literatura está el trabajo de (Jangir et al., 2014), donde se realiza el procesamiento de la señal de la voz para generar un sistema de identificación de voz, el cual se utiliza la correlación de espectros y el método de normalización en tiempo real. Utilizando el espectrograma obtienen información como intensidad, frecuencia y magnitud en forma gráfica, la cual se vuelve útil para la identificación de voz. Dentro de su investigación llegan a la resolución de que el método de correlación tiene mayor tasa de reconocimiento y mediante el uso del método de normalización que proponen logran superar la limitación del proceso estático. El método de normalización entrega un rendimiento dinámico al momento de reconocer y se puede mejorar reduciendo el ruido mediante un proceso de filtrado durante la grabación del sonido con el micrófono.

3.2. Trabajos utilizando Redes Neuronales Profundas

En el trabajo de (Variani et al., 2014) proponen un sistema de verificación de voz dependiente de texto, en el cual utilizan una DNN supervisada para la extracción de características a nivel cuadro. De esta manera, cada cuadro se va apilando en un vector de izquierda a derecha, correspondiendo al número de voces de entrenamiento por cada cuadro. Después de entrenar satisfactoriamente, las activaciones de salida acumuladas de la última capa oculta pasan a ser la nueva representación de voz, ahora cada fotograma de un texto dado que pertenezca a un nuevo hablante. Paso siguiente, se calcula la activación de salida de la última capa oculta utilizando propagación hacia delante en la DNN entrenada, y al final se hace la acumulación de esas activaciones para formar una representación compacta de ese hablante, el vector d. La verificación al final se hace calculando la distancia del coseno entre el vector d de la prueba y del hablante a evaluar, así utilizando un umbral para tomar la decisión.

En el artículo (Plchot et al., 2016) se utiliza una modificación de la DNN, donde se emplea la combinación de un autocodificador con DNN, esto con la finalidad de mejorar el audio, donde la función principal del autocodificador es la eliminación del ruido y la reverberación.

Los autores en (Boles & Rad, 2017) en su trabajo describen un sistema de identificación por voz independiente de texto, donde mediante el uso de Coeficientes Cepstrales de Frecuencia de Mel se evaluó el sistema. Además, se utiliza la máquina de soporte de vectores, la cual fue entrenada y probada con dos distintos conjuntos de datos, LibriSpeech siendo uno de dominio público y el segundo fue el utilizar audios grabados dentro de una casa. Contando así con dos distintas contribuciones, la primera es el uso de la Maquia de soporte de vectores con un extractor de características de Coeficientes Cepstrales de la Frecuencia de Mel y segundo un audio único es entrenado en la red neuronal dentro de un servidor “Bare metal” el cual logra un entrenamiento rápido y de igual modo una clasificación rápida.

En el trabajo de (Bae et al., 2016) mediante el uso del aprendizaje profundo y usando los Coeficientes Cepstrales de Frecuencias de Mel Adaptativa, se propone un método de reconocimiento de voz mejorado. Donde la utilización de MFCC adaptativo para la extracción de datos de una señal de audio son más eficientes, ya que a diferencia de los otros existentes, este no deteriora el audio. El filtro MFCC adaptativo está construido de una forma compacta

en el área de densidad de datos para así mitigar la perdida de características del audio. Obteniendo así una mejor tasa de reconocimiento. Además del sistema propuesto con aprendizaje profundo, permite utilizar el reconocimiento de voz sin base de datos, permitiendo utilizar dispositivos de poca capacidad de almacenamiento.

3.3. Trabajos enfocados en Redes Neuronales Convolucionales

Un modelo enfocado en CNN propuesto en (Lukic et al., 2016) está diseñado para optimizar el proceso de identificación de voz, basado en el dataset TIMIT. Para el preprocesamiento de datos se utilizan espectrogramas para mejorar las fuentes acústicas. Esta CNN contiene un gran numero de capas de convolución aplicando varios filtros a diferentes secciones locales de entrada, a la cual esa capa le sigue una capa de agrupación máxima. Esta emite una versión de más baja resolución de las activaciones de la capa de convolución eliminando la activación total del filtro. Por último, las capas totalmente conectadas acoplan todas las salidas de la última capa de max-pooling para clasificar las voces.

Dentro de la categoría de las CNN existen las redes convolucionales 3D, donde el kernel encargado de la convolución se mueve en 3 direcciones. En el trabajo (Torfi et al., 2018) se propone la utilización de las 3D-CNN, donde estas tres dimensiones, además del dominio de la frecuencia como en otros trabajos, también se enfoca en el dominio del tiempo y el tamaño de enunciados que existe en un audio, con la finalidad de construir un modelo más robusto para la problemática de los cambios que existen en la voz.

Además de los modelos GMM, HMM o los de aprendizaje profundo, esta el modelo basado en el vector i para la detección automática de voz en línea. El vector i extrae de cada voz un vector de longitud fija con la finalidad de crear una plantilla. En el trabajo de (Kalimoldayev et al., 2019) se propone un modelo basado en el vector i, el cual es un espacio de baja dimensión y llegan a la conclusión que los vectores i son las funciones más eficaces para la verificación de personas que no dependen de un texto. Durante el estudio, ellos analizan como el parámetro del vector i, el tamaño del *Universal Background Model* (UBM) y la dimensión del vector i, afectan la precisión de la identificación por voz.

En el trabajo de (Hendryli & Herwindiati, 2020) desarrollan un sistema de autentificación por voz en el cual se implementa la verificación de la persona mediante un sistema de contraseña de un solo uso (OTP por sus siglas en ingles) y un modelo de reconocimiento de voz. El modelo del reconocimiento de voz se encarga de detectar las expresiones del usuario de los dígitos aleatorios del OTP. El modelo de verificación se encarga de verificar la similitud de la voz del usuario. Para el desarrollo de este sistema, se utiliza una red siamesa y una red de memoria a corto plazo, estas serán las encargadas de reconocer y verificar mediante la utilización de los Coeficientes Cepstrales de Frecuencias de Mel. Este sistema obtiene gran precisión en el reconocimiento de voz, pero el modelo de verificación no alcanza resultados satisfactorios.

Tabla del estado del arte

Tabla 3.1: Descripción de los sistemas de reconocimiento de voz del estado del arte

Autor y año	Base de datos utilizada	No. De hablantes	Entrada	Modelo	Desempeño
Jangir et al., 2014	independiente	5	Espectrograma	HMM Distribución multiespacio	AC: 94
Variani et al., 2014	-	646	Características energéticas del marco	DNN	EER : 2.00(Por 20 expresiones)
Das & Nahar, 2016	independiente	NA	MFCC	HMM	AC: 90
Bae et al., 2016	-	10000	MFCC Adaptativo	DNN	-
Lukic et al., 2016	TIMIT	630	Espectrograma de datos de voz	CNN	AC: 97
Plchot et al., 2016	Fisher corpora PRISM Switch Board SRE	13916 1991 2740	MFCC, PNCC	DNN auto encoder	-
Boles & Rad, 2017	LibriSpeech		MFCC	SVM	
Torfi et al., 2018	WVU-Multimodal 2013	1083	Marco de la MFEC	3D-CNN	EER: 21.1
Tamoto & Itou, 2019	AWA-LTR	40	MFCC	Posteriograma GMM	
Kalimoldayev et al., 2019	-		vector i	SVM	-
Dhakal et al., 2019	ELSDSR	22	Gabor+CNN+Estadistico	SVM RF DNN	AC: 98.07 AC: 99.41 AC: 98.14
Jahangir et al., 2020	LibriSpeech	50 Hombres 50 Mujeres	MFCCT	MLDNN	AC: 92.9
Hendryli & Herwindiati, 2020	independiente	905	MFCC	CNN Siamesa	AC: 70 %

Capítulo 4

Metodología

Durante la etapa del desarrollo del sistema de identificación de voz, fue necesario utilizar hardware como parte del proceso de desarrollo del sistema. De esta manera, a continuación se dará una descripción del hardware utilizado y posteriormente se explicará el proceso de desarrollo para el sistema a nivel computacional.

4.1. Hardware utilizado

Como todo sistema biométrico, es necesario contar con un dispositivo que sea capaz de capturar el rasgo biométrico a utilizar como método de autentificación. Para el desarrollo del sistema es necesario un micrófono con propósito de contar con el dispositivo de entrada que capture la voz para el sistema y tener una entrada controlada. Es decir, que no existan variaciones en la voz utilizando distintos micrófonos para las pruebas. El micrófono utilizado es un HiperX QuadCast, el cual cuenta con modulador de sensibilidad físico para tener control de las ganancias de los audios para lograr reducir los ruidos ambientales que afectan a gran escala el reconocimiento de voz.

Las características del micrófono se muestra en la Tabla 4.1, de esta manera el proceso de registro y de autentificación tendrán las mismas características en el audio capturado. Con el mismo propósito, el dispositivo fue utilizado para la creación del *dataset* de distintas frases, para las pruebas del sistema y así no exista variabilidad entre los audios capturados.

Tabla 4.1: Características de micrófono HyperX QuadCast

Características	Valores
Velocidad de muestreo/bits	48kHz/16-bit
Patrones polares	Estéreo, omnidireccional, cardiode, bidireccional
Respuesta de frecuencia	20Hz-20kHz
Sensibilidad	-36dB(1V/Pa a 1kHz)

El procesamiento del software se realiza mediante una PC Servidor local, la cual cuenta con distintas características para realizar las tareas de identificación de voz sin ningún problema. Las características se muestran en la Tabla 4.2

Tabla 4.2: Características del servidor

Nombre	Características
Sistema Operativo	Windows 11
RAM	16 GB
Almacenamiento	500 GB SSD
Procesador	Intel Core i7 9° Gen
Tarjeta Gráfica	Nvidia RTX 2060

4.2. Desarrollo del sistema de identificación por voz

A partir de conocer la estructura general de un sistema biométrico, el cual se muestra en la Figura 2.20, es posible trabajar tanto como la metodología tradicional de Top-Down, como las metodologías ágiles, específicamente un desarrollo iterativo e incremental. Se optó por este último debido a las características que brinda, donde el proceso de diseño se puede dividir en pequeños proyectos y fáciles de manejar.

4.2.1. Análisis de requisitos

Durante la investigación de los sistemas biométricos se obtuvo conocimiento sobre el funcionamiento de estos. Como se mencionó anteriormente, un sistema biométrico cuenta con ciertas características generales que se consideran como plantilla para partir en el desarrollo del sistema. De esta manera se plantearon los requisitos tanto funcionales como los requisitos de calidad que debe cumplir el sistema de identificación por voz. Una vez obtenidos la especificación de requerimientos, se podrá continuar con el diseño de la arquitectura.

4.2.1.1. Requisitos Funcionales

Estos requisitos describen la funcionalidad del sistema a desarrollar, de esta manera, los requisitos funcionales planteados para que el sistema de identificación de voz cumpla su propósito, se generaron a partir de como fue entrenada la red neuronal a utilizar, de que manera se va almacenar los usuarios permitidos a autentificarse y de que manera se van autenficar los usuarios. A partir de esto, se generaron los siguientes requisitos funcionales:

- RF01.- El sistema deberá registrar en una base de datos una carpeta con un identificador de usuario con el audio de registro, junto con un vector de características obtenido por una red neuronal.
- RF02.- El sistema deberá de utilizar un micrófono como único método de entrada de audio y tener un formato WAV, Canal Mono, 16 Bits de resolución a 16kHz de frecuencia.
- RF03.- El sistema deberá capturar audios de 3 a 10 segundos máximo y obtener su vector de caracteristicas para compararlo con los usuarios permitidos.
- RF04.- El sistema deberá de autentificar personas mediante la voz, tomando la base de datos de personas permitidas.
- RF05.- El sistema deberá desplegar un mensaje de autentificación de la persona, mostrando el ID de la persona (si es aceptada) o indicar que no se reconoce al usuario.

4.2.1.2. Requisitos de Calidad

El trabajo desarrollado por (Rui & Yan, 2018), el cual proporciona una lista de evaluaciones de calidad para los sistemas biométricos. Los cuales, son punto de partida a considerar como requisitos de calidad, y se describen Tabla 2.3, se tomaran solo 3 de los 5 puntos: Exactitud, Eficiencia y Usabilidad, debido a que los otros dos no fueron implementados durante el desarrollo de la tesis. Por otra parte, mediante la creación de una base de datos propia de distintas frases por voz, se evaluaron estos puntos para la etapa de resultados.

4.2.2. Actores

Los actores representan todo lo que está interactuando con el sistema para que funcione de manera correcta. El sistema de identificación por voz, cuenta con solo dos actores donde ambos son usuarios. Los actores están definidos por administrador y usuario.

- Administrador:

  Es el encargado de registrar a los usuarios autorizados dentro del sistema. El administrador también es un usuario dentro del sistema.

- Usuario:

  Es la persona que se quiere autentificar mediante la voz, la cual debe proporcionar una frase al micrófono.

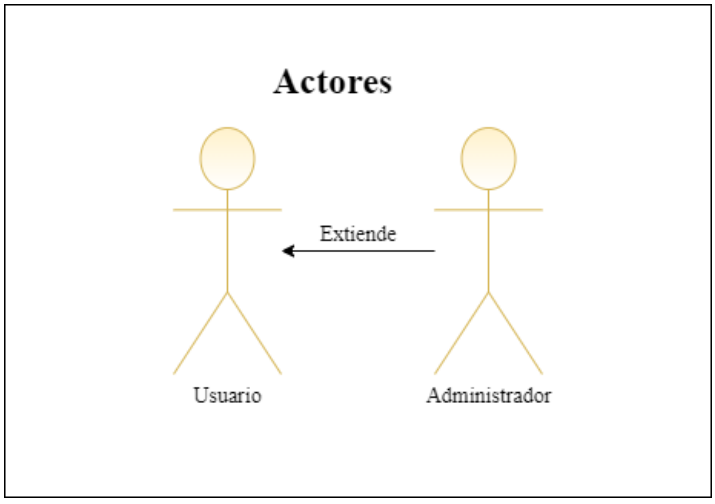

Figura 4.1: Actores del sistema

4.2.3. Casos de uso

Mediante los casos de uso se mapean los requisitos funcionales descritos anteriormente, para así tener un diagrama de interacción de los actores con el sistema. El identificador por voz cuenta con dos casos de uso principales, los cuales se observan en el diagrama de la Figura 4.2.

- CU00.- Registrar usuario nuevo:

  Es la representación del registro de un usuario al sistema almacenando sus respectivos archivos y un ID.

- CU01.-

  Identificador por voz: El usuario o el administrador ingresan su voz diciendo una frase en el micrófono y se hace el proceso de identificación dentro del sistema.

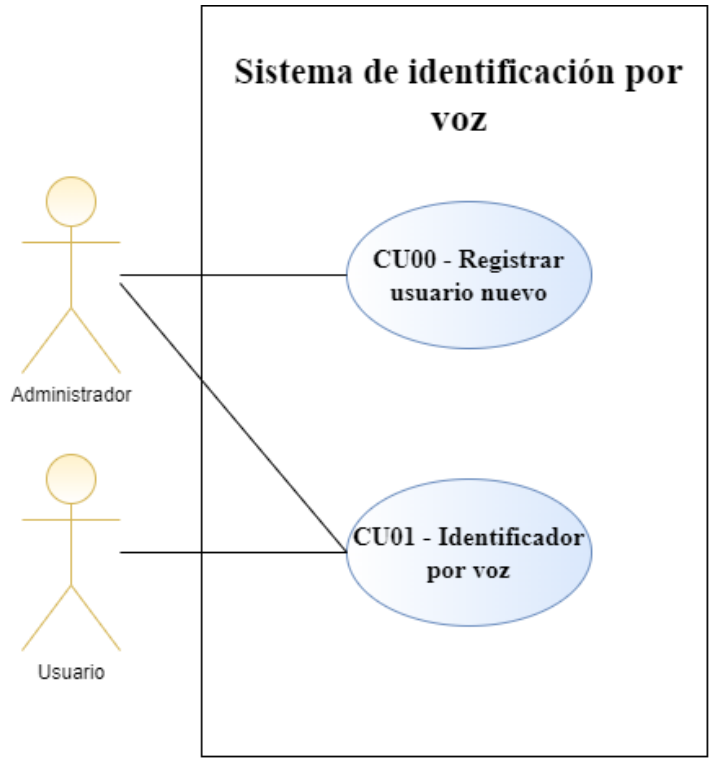

Figura 4.2: Diagrama Casos de Uso

4.3. Diagrama de Contexto

Mediante el uso del diagrama de contexto se observa el flujo de la información para conocer como interacciona el sistema con el exterior. En la Figura 4.3 se muestran los que existen dentro del diagrama de contexto, donde se observa quien usa, que usa y de quien depende el sistema. De tal manera que, el usuario y el administrador utilizan el sistema de identificación por voz. Para esto el sistema usa el micrófono como método de entrada y por último el sistema depende de la base de datos para almacenar a los usuarios registrados.

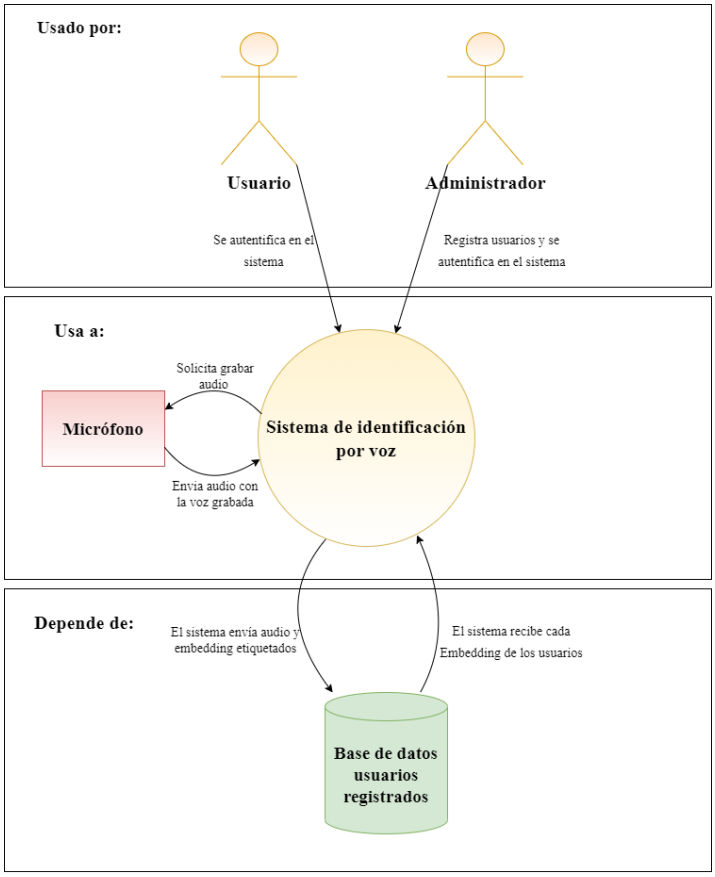

Figura 4.3: Diagrama de Contexto

- Usuario:

  Puede ser usuario o administrador, donde la función adicional que cumple el administrador es registrar a los usuarios en el sistema. Ambos utilizan la identificación de voz.

- Micrófono:

  Dispositivo encargado de capturar la entrada del usuario, es decir, una frase dicha con su voz, siendo así el sensor que capta el audio para ser procesado.

- Base de datos de usuarios registrados:

  Encargada de almacenar en un directorio con el ID y dentro de ellas el audio de la voz y el vector de características extraído.

  Almacena los datos de usuario con el ID, audio de la frase de registro y su vector de características extraído.

4.4. Arquetipos

La utilización de los arquetipos proporciona la representación más abstracta de las pequeñas entidades dentro del sistema para así describir la mayor parte de como se comporta el sistema de identificación por voz. Dentro del sistema desarrollado obtuvo cinco arquetipos en total para la representación del sistema, los cuales se describen a continuación.

- Usuario:

  Es la representación de la persona que va a proporcionar su rasgo biométrico para el sistema, la cual va a ser registrada en el sistema para ser identificada.

- Administrador:

  Tiene la misma función del usuario, con la diferencia de tener la capacidad de registrar a las personas que están autorizadas a ser identificadas dentro del sistema.

- Identificador:

  Es la abstracción de la función principal del sistema, la cual es asignar un valor ID a la persona que quiere identificarse, siempre y cuando este registrado.

- Voz:

Para ser identificada una persona es necesario contar con este rasgo biométrico. La voz es la representación abstracta del dato proporcionado por el usuario al momento de identificarse o registrarse.

- Vector de características:

  Es la representación abstracta de los números o características más representativas de la voz, donde se obtiene mediante un preprocesamiento.

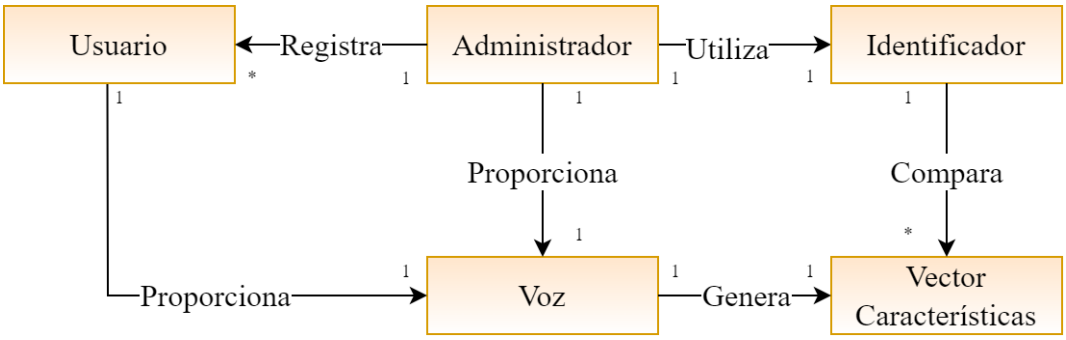

Figura 4.4: Diagrama de Arquetipos

En la Figura 4.4 se aprecia gráficamente como es que se relacionan todos los arquetipos descritos anteriormente.

4.5. Arquitectura

De manera general, la arquitectura planteada para el proyecto está enfocada en el modelo tubos y filtros, debido al proceso de comunicación entre cada bloque del sistema. Partiendo de un modelo general de un sistema biométrico, desde la etapa de inscripción de usuario hasta la etapa del comparador del dato biométrico, se desarrolla el sistema de identificación de voz descrito en la figura 4.5.

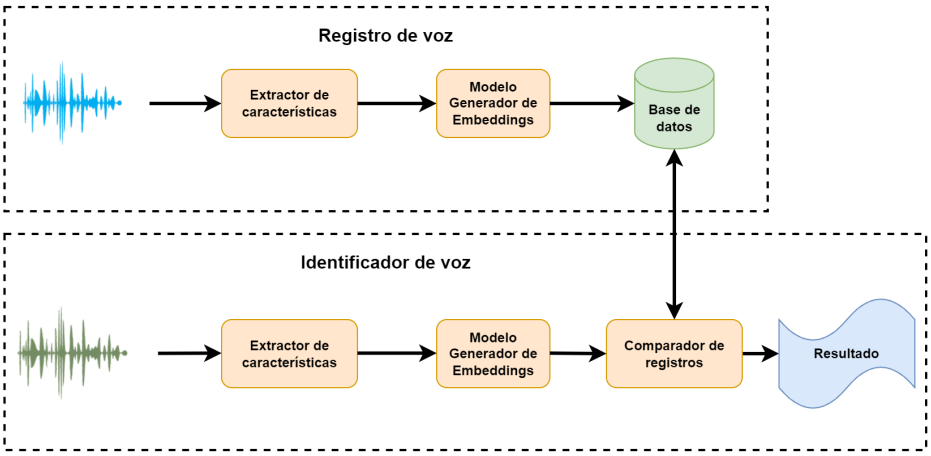

Figura 4.5: Arquitectura general del sistema de identificación por voz

El sistema de identificación por voz se separa en dos etapas, la de registro de usuario y la de identificador de voz. Las cuales a continuación se describirán con sus procesos a detalle.

Etapa de registro de usuario:

En la figura 4.6 se observa el proceso que se lleva para registrar un usuario de una manera más detallada.

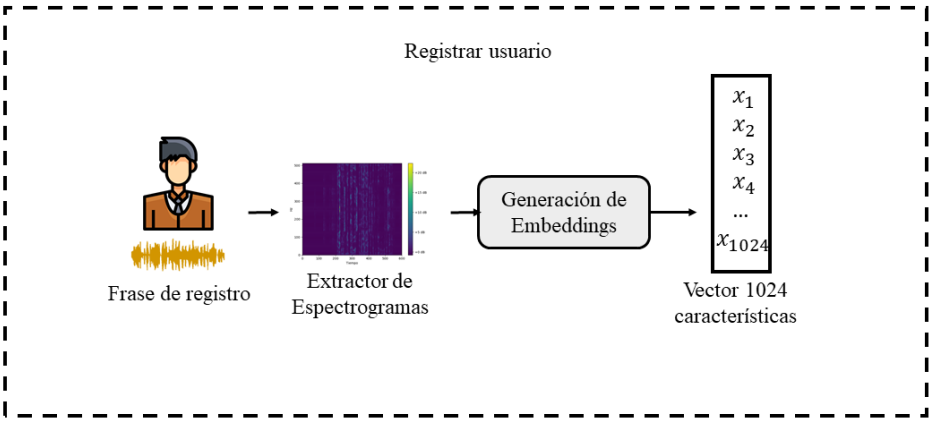

Figura 4.6: Diagrama de etapa de registro de voz

- **Frase de registro:**

  La primera parte de la etapa de registro es la entrada de audio llegando en crudo, con un formato tipo WAV, a una frecuencia de 16kHz, en canal mono y una tasa de muestreo de 16-Bits. Debido a que el modelo fue entrenado con las características descritas

anteriormente, es necesario que tanto como el registro y la etapa de identificación de voz, la entrada de audios cumpla con estas características. Para registrar un usuario es estableció una frase en concreto que dura aproximadamente 5 segundos. La frase utilizada fue “El reconocimiento de voz es la llave de acceso al futuro como contraseña”, con la finalidad de posteriormente hacer pruebas con la misma frase de registro y con frases distintas para verificar el impacto que tiene al momento de identificar.

- **Extractor de espectrogramas:**

  El tipo de extractor de característica seleccionado para el sistema de identificación de voz se denomina espectrograma, esto debido a que una de las principales características que hacen única a la voz es el timbre. Debido a que el timbre trabaja con las frecuencias, el espectrograma se encarga de extraer estas frecuencias que se encuentran en la voz. A partir del audio en “crudo” previo a la extracción del espectrograma se realiza un preprocesamiento utilizando la *Transformada Rápida de Fourier*, la cual pasa del dominio del tiempo al dominio de la frecuencia. Una vez transformado el audio al dominio de la frecuencia se extrae el espectrograma del audio, el cual se representa en la figura 4.7.

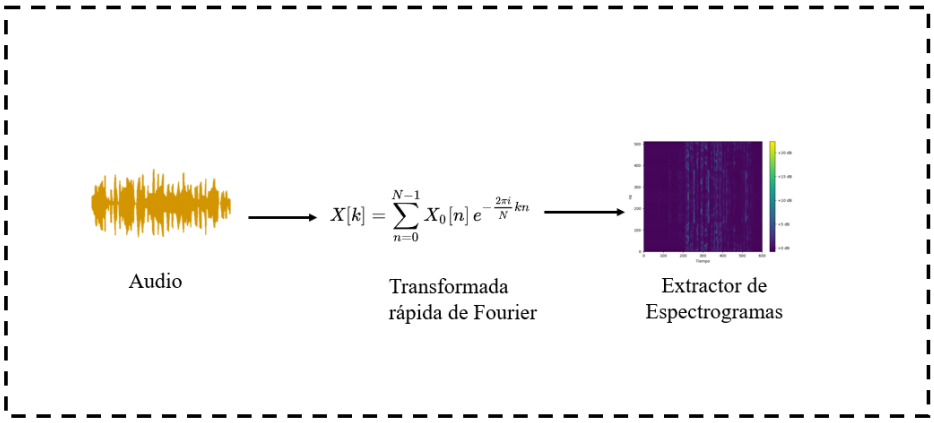

Figura 4.7: Extracción de Espectrograma

- **Generación de embeddings:**

  Para la etapa de extracción del vector de características, se utilizó el modelo de red neuronal llamado VGGVox, desarrollado en el trabajo (Nagrani et al., 2017). Esta red

fue desarrollada a partir de la red neuronal VGG-M (Chatfield et al., 2014), que es una red neuronal convolucional enfocado en la clasificación de imágenes. Con los cambios realizados en la arquitectura de la red neuronal de la VGG-M a la VGGVox se redujo de 319 millones de parámetros a entrenar a 67 millones para así evitar el sobreentrenamiento. En la figura 4.8 se observa la estructura general de la Red neuronal Convolucional VGGVox. El modelo VGGVox fue adaptado para una entrada de espectrogramas, se cambió de la VGG-M la capa *fully Connected 6* por dos capas donde la capa "*fc6*" utilizada para los audios en el dominio del tiempo y la capa "*average pool*" el cual utiliza un kernel de 1xn donde n varía dependiendo del tamaño del tiempo.

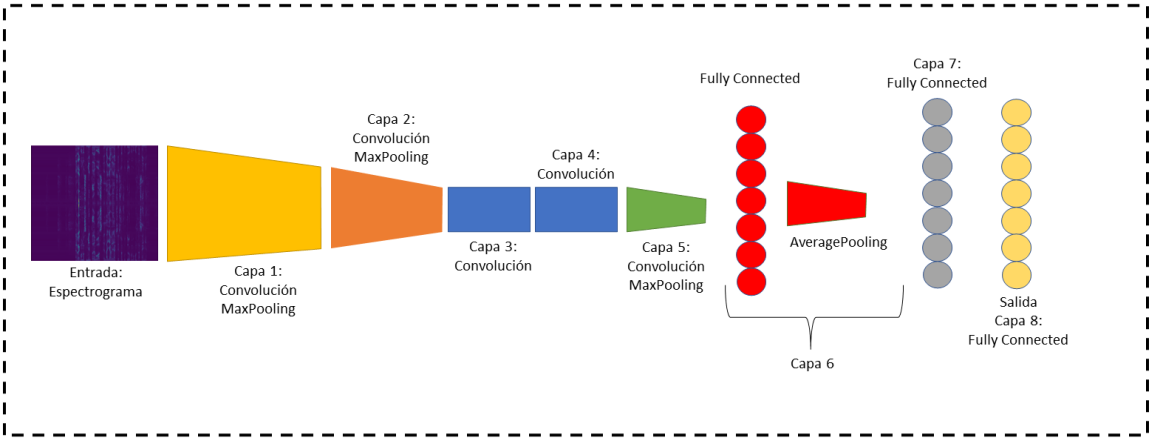

Figura 4.8: Red Neuronal Convolucional VGGVox

En la Tabla 4.3 se observa como está estructurada la arquitectura y las capas que fueron modificadas *fc6* y *apool6*, además se aprecia su kernel, dimensión de filtro, el tamaño de paso del kernel y el tamaño de datos.

Tabla 4.3: Arquitectura VGGVox

Capa	Kernel	Dimensión de filtro	Filtros	Stride	Tamaño de datos
conv1	7x7	1	96	2x2	256x148
mpool1	3x3	-	-	2x2	126x73
conv2	5x5	96	256	2x2	62x36
mpool2	3x3	-	-	2x2	30x17
conv3	3x3	256	384	1x1	30x17
conv4	3x3	384	256	1x1	30x17
conv5	3x3	256	256	1x1	30x17
mpool5	5x3	-	-	3x2	9x8
fc6	9x1	256	4096	1x1	1x8
apool6	1xn	-	-	1x1	1x1
fc7	1x1	4096	1024	1x1	1x1
fc8	1x1	1024	1251	1x1	1x1

- **Vector de característica:**

  El vector de características denominado también *embedding* tiene un tamaño de 1024, donde surge a partir de los datos de entrada de una matriz de 512 x n donde n varía dependiendo los segundos del audio.

Etapa de identificación de usuario:

La etapa de identificación se observa gráficamente en la figura 4.9, donde se tiene un proceso similar al registro desde el ingreso del audio hasta la obtención del "*embedding*". Una vez se obtiene el vector de características llega la etapa del comparador, el cual se encarga de utilizar el vector de características del audio de entrada y encontrar a cuál voz es más similar en la base de datos.

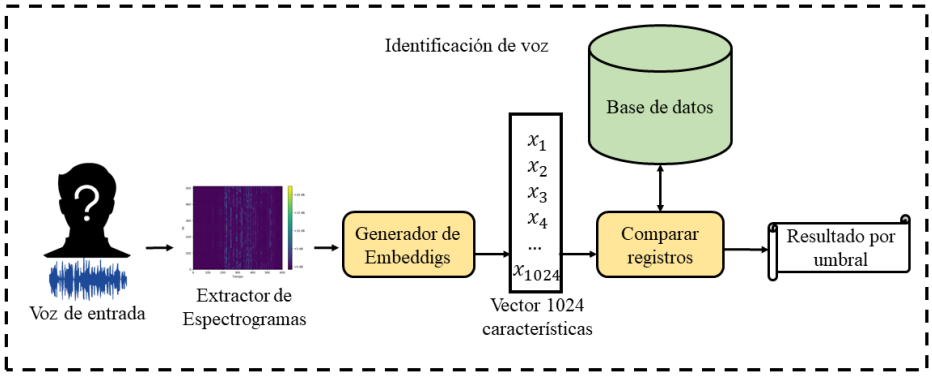

Figura 4.9: Diagrama de etapa de identificación de voz

- **Voz de entrada:**

  Esta etapa es el inicio del proceso de identificación por voz, en la cual mediante una frase corta de un usuario pasa por la extracción de espectrograma y se obtiene el *embedding*. Para la etapa de pruebas se obtuvo una base de datos de alrededor 128 personas, con 6 audios de prueba cada uno, obteniendo así 768. La mitad de los audios fue utilizando la frase de registro "el reconocimiento de voz es la llave de acceso al futuro como contraseña" y la otra mitad fue utilizando frases distintas. La finalidad de grabar distintas frases fue el conocer si existe variación en la identificación al momento de decir la misma frase a una distinta.

- **Comparar registros:**

  Al tener el vector de características de la voz de entrada llega la etapa identificar si la voz coincide con alguna de las registradas en la base de datos. En esta etapa, al ser vectores de mismo tamaño de características y solo trabajar en un mismo plano (frecuencias) mediante la utilización de la ecuación 4.1 es considerada una de las métricas más utilizadas y de menor costo computacional al momento de comparar similitud entre vectores.

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.1}$$

  Donde y representa el vector de entrada y x representa las voces registradas, esta última iterándose para medirse una a una con la voz a identificar y se guarda cada distancia en

un vector para la siguiente etapa. La finalidad de la medición de la distancia euclidiana es que entre más cercano a 0 sea el valor obtenido, significa que más acertada es la identificación.

En la Figura 4.10 se observa la forma de como se comparan los registros de usuarios contra el vector de entrada, donde el vector de entrada tiene un ID desconocido y se calcula la similitud mediante el uso de la distancia euclidiana descrita en la Ecuación 4.1, comparándolos también de 1 a N personas.

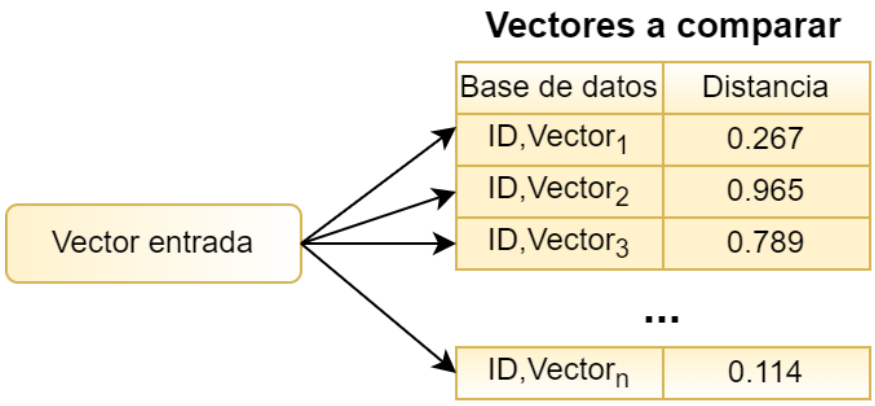

Figura 4.10: Comparación de usuarios registrados vs vector de entrada

- **Resultado por umbral:**

  Al obtener el vector de distancias es necesario determinar si esa persona identificada es la correcta o no, es por eso que mediante un umbral se determina la identificación de la persona. Al ir aumentando el tamaño de personas registradas, este umbral debe ir variando, es decir, ir disminuyendo este parámetro conforme va aumentando el registro de personas para así tener mayor tasa de precisión y evitar la identificación errónea. Debido a que se esta utilizando la medida de distancia euclidiana para identificar a la persona, entre menor sea el umbral, significa que es más similar es la voz registrada a la de entrada al sistema. El umbral va en función a los valores que se obtienen con la distancia euclidiana.

  Existen otras métricas que ayudan a conocer que tan fiable es el sistema de identificación. La exactitud que es interpretada por la tasa de porcentaje de la cantidad de

predicciones positivas que fueron correctas, y se obtienen con la Ecuación 4.2.

$$exactitud = \frac{(VP + VN)}{(VP + FP + FN + VN)} \tag{4.2}$$

Ahora bien, para obtener el valor de precisión, el cual representa el porcentaje de casos positivos detectados, donde este porcentaje indica que tan fiable es el valor detectado como positivo, se define con la Ecuación 4.3.

$$precision = \frac{VP}{(VP + FP)} \tag{4.3}$$

4.6. Creación de *dataset* de voces con frases distintas

Para la etapa de pruebas y resultados, era necesario contar con un *dataset* de voces para someter al sistema desarrollado anteriormente. De tal manera que, se desarrolló un *dataset* el cual cumpliera con el objetivo de las pruebas a realizar. El desarrollo del *dataset* se enfocó en obtener audios con distintas frases de diversas personas. En total se recopilaron 768 audios, de 128 personas, donde la mitad fueron diciendo la misma frase con la que se registraron y la otra mitad era diciendo una frase distinta proporcionada por una lista.

En la Tabla 4.4 se muestran las frases utilizadas, donde la frase de registro de los usuarios, fue la asignada para grabar la mitad de los audios, los cuales fueron un total de 384. Para grabar los audios de frases distintas, se seleccionaba aleatoriamente por el usuario de una lista proporcionada, las cuales están listadas en la misma tabla.

Frase de registro	Frases distintas
El reconocimiento de voz es la llave de acceso al futuro como contraseña	La vida comienza al final de tu zona de confort El éxito no se mide en el dinero, sino en la diferencia que marcas en las personas La lógica te llevará desde A hasta B. La imaginación te llevará a cualquier parte Solo se vive una vez. Pero si lo haces bien, una vez es suficiente. Los dos días más importantes de tu vida son el día en que naciste y el día en que encontraste el por qué.

Tabla 4.4: Frases para *dataset* diseñado

Para del desarrollo del *dataset*, en la Figura 4.11 se observa el proceso para almacenar la grabación, los cuales se describirán brevemente a continuación:

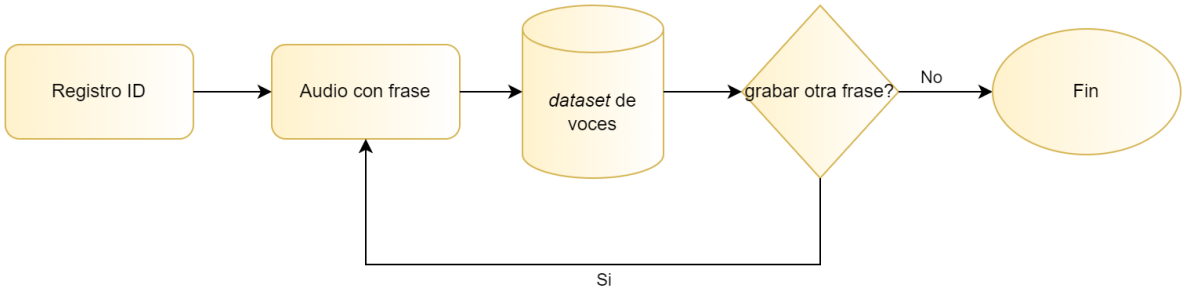

Figura 4.11: Diagrama de creación de *dataset* de voces

- **Registro ID**:

  Durante la primera etapa de la creación del *dataset*, se introduce una etiqueta o ID para el registro y dirección donde se va a almacenar cada uno de los audios de distintas frases.

- **Audio con frase:**

  Para esta etapa, la persona a registrar ingresa la frase mediante el micrófono. Los cuales estos audios mantienen el mismo formato de trabajo que se ha utilizado para el desa-

rrollo del sistema de identificación por voz. El formato es WAV, a 16 Bits de resolución y una frecuencia de 16kHz.

- ***dataset* de voces:**

  En el *dataset* de voces, se crea la dirección con el ID que se ingresó y la grabación de la frase.

- **Grabar otra frase:**

  Una vez se crea el directorio y se almacena el primer audio de la persona, se realiza una toma de decisión donde se da la posibilidad de almacenar más audios en esa misma dirección. Se realiza 5 veces más este proceso para introducir las distintas frases en una misma etiqueta.

En la Figura 4.12 se observa el proceso de grabación de una persona, donde se utilizó el mismo micrófono y PC, donde se desarrolló el sistema de identificación por voz. Por otro lado, para tener un *dataset* más controlado, se realizó en un área cerrada para evitar lo más posible los ruidos ambientales.

Figura 4.12: Grabación de *dataset*

Capítulo 5

Resultados y análisis

Al concluir con el desarrollo del sistema de identificación de voz, fue necesario someterlo a distintas pruebas para comprobar la confiabilidad del sistema. Mediante uso de herramientas estadísticas como la matriz de confusión se realizó el análisis del sistema. Además, se usaron las métricas de desempeño para clasificación y adicionalmente las medidas para sistemas biométricos.

Resultados

Para la obtención de resultados, fue necesario la preparación de las pruebas, donde se hizo el registro del usuario para posteriormente implementar el *dataset* desarrollado. Para el desarrollo del *dataset* descrito en la Figura 4.11, la población que participó fueron estudiantes del Instituto Tecnológico de Culiacán. Al rededor de 128 estudiantes entre 19 a 21 años, sin considerar el sexo y generando cada uno seis audios en total.

De las 128 personas del *dataset*, se registraron 90 y 38 permanecieron fuera del sistema. Posteriormente, se sometieron los 768 audios de prueba para obtener los resultados. En la Tabla 5.1 mediante la utilización de las 4 opciones que ofrece la matriz de confusión, se exploraron distintos umbrales para la toma de decisión de similitud de audios. Se obtuvieron los Verdaderos Positivos (VP) que son aquellos que el sistema detecta como verdaderos y su valor real es verdadero. Los Falsos Positivos (FP) se interpretan como los audios aceptados como una persona distinta. Los Verdaderos Negativos (VN) estos resultados se obtienen de personas que no están en el sistema y que realmente los rechaza. Por último, los Falsos

Negativos (FN) que son aquellos audios que si debieron asignarle un valor verdadero, pero que fueron rechazados por el sistema.

Tabla 5.1: Matriz de confusión con audios diciendo la misma frase de registro

Umbral	Verdaderos positivos	Falsos positivos	Verdaderos negativos	Falsos negativos
0.1	136	9	85	154
0.12	175	20	76	113
0.14	210	29	72	73
0.16	237	40	63	44
0.18	248	51	55	30
0.2	254	68	44	18
0.22	259	77	37	11
0.24	262	86	30	6
0.26	264	93	26	1
0.28	265	101	18	0
0.3	265	106	13	0

A partir de la Tabla 5.1 de los 4 puntos de la matriz de confusión, se determinaron las métricas por cada umbral de exactitud, precisión, F1, FAR y FRR las cuales se pueden observar en la Tabla 5.2. Estas dos últimas métricas son utilizadas para evaluar los sistemas biométricos. De tal manera que, los valores de FAR indican que tan preciso es el sistema al momento de autentificar a un usuario y los del FRR indican que tan propenso esta el sistema en denegar la autentificación a una persona que si pertenece al sistema. En ambas métricas entre más cercano este al 0 % indica que es mejor en esa evaluación.

Dado un ejemplo, en la Tabla 5.2 el umbral resaltado en 0.3, indica que la exactitud del sistema para determinar un positivo, sin importar si es falso o verdadero es del 72.40 % y la precisión obtenida dice que el 72.43 % el sistema clasifica bien como verdadero. Ahora bien, el FAR de 89.08 % indica que el sistema tiene un alto porcentaje en aceptar falsos positivos, o permitir usuarios no autorizados. Por ultimo, el FRR, se observa de 0 % debido a que este valor describe que tantas personas rechaza. El caso ideal sería tener un 100 % de exactitud y precisión y 0 % en FAR y FRR.

Tabla 5.2: Métricas obtenidas de la matriz de confusión de audios diciendo la misma frase de registro

Umbral	Exactitud	Precisión	F1	FAR	FRR
0.1	57.55 %	93.79 %	62.53 %	9.57 %	53.10 %
0.12	65.36 %	89.74 %	72.46 %	20.83 %	39.24 %
0.14	73.44 %	87.87 %	80.46 %	28.71 %	25.80 %
0.16	78.13 %	85.56 %	84.95 %	38.83 %	15.66 %
0.18	78.91 %	82.94 %	85.96 %	48.11 %	10.79 %
0.2	77.60 %	78.88 %	85.52 %	60.71 %	6.62 %
0.22	77.08 %	77.08 %	85.48 %	67.54 %	4.07 %
0.24	76.04 %	75.29 %	85.06 %	74.14 %	2.24 %
0.26	75.52 %	73.95 %	84.89 %	78.15 %	0.38 %
0.28	73.70 %	72.40 %	83.99 %	84.87 %	0.00 %
0.3	*72.40 %*	*71.43 %*	*83.33 %*	*89.08 %*	*0.00 %*

En la Tabla 5.3 se muestra los resultados de la matriz de confusión para frases distintas con el mismo tamaño de audios. Además, que en la Tabla 5.4 se muestran las métricas obtenidas.

Tabla 5.3: Matriz de confusión con audios diciendo distinta frase a la de registro

Umbral	Verdaderos positivos	Falsos positivos	Verdaderos negativos	Falsos negativos
0.1	60	5	90	227
0.12	96	21	86	179
0.14	125	41	78	138
0.16	149	64	69	100
0.18	173	93	55	61
0.2	189	112	47	34
0.22	196	128	34	24
0.24	200	141	26	15
0.26	205	149	21	7
0.28	207	156	15	4
0.3	208	160	11	3

Al comparar resultados entre la Tabla 5.1 y 5.3 de matriz de confusión con la misma frase y frase distinta, se puede observar que existe una diferencia de aproximadamente un 20 % entre los Verdaderos positivos. Es decir, que al utilizar la misma frase existe mayor exactitud a la hora de reconocer correctamente a la persona. Por otro lado, un punto importante de la matriz de confusión en los sistemas biométricos son los falsos positivos, ya que estos determinan cuantas personas identificó erróneamente. Al comparar ambas pruebas conforme va creciendo el umbral en el sistema empieza aumentar los falsos positivos, teniendo que la diferencia entre el umbral más bajo, no hay una diferencia significativa. Ahora, al utilizar un

umbral alto, como es el caso de 0.3 la diferencia entre ellos es casi de 15 %. Por lo tanto, no es conveniente en los sistemas biométricos una tasa alta en falsos positivos, ya que esto implica asignar erróneamente un valor correcto a una persona.

Tabla 5.4: Métricas obtenidas de la matriz de confusión de audios diciendo distinta frase a la de registro

Umbral	Exactitud	Precisión	F1	FAR	FRR
0.1	39.27 %	92.31 %	34.09 %	5.26 %	79.09 %
0.12	47.64 %	82.05 %	48.98 %	19.63 %	65.09 %
0.14	53.14 %	75.30 %	58.28 %	34.45 %	52.47 %
0.16	57.07 %	69.95 %	64.50 %	48.12 %	40.16 %
0.18	59.69 %	65.04 %	69.20 %	62.84 %	26.07 %
0.2	61.78 %	62.79 %	72.14 %	70.44 %	15.25 %
0.22	60.21 %	60.49 %	72.06 %	79.01 %	10.91 %
0.24	59.16 %	58.65 %	71.94 %	84.43 %	6.98 %
0.26	59.16 %	57.91 %	72.44 %	87.65 %	3.30 %
0.28	58.12 %	57.02 %	72.13 %	91.23 %	1.90 %
0.3	57.33 %	56.52 %	71.85 %	93.57 %	1.42 %

Las métricas de exactitud y precisión en ambas pruebas de las Tablas 5.2 y 5.4, se puede observar que el sistema es preciso con umbrales bajos, esto debido a que, entre más bajo sea el umbral, más cerca está de ser similar la voz de entrada con la que se tiene registrada.

La precisión en este caso indica la fiabilidad o certeza con la que el sistema está clasificando el valor positivo. Es decir, para evitar que el sistema se confunda por la similitud en las personas en su timbre de voz, es mejor utilizar umbrales bajos para así asegurar mayor precisión a la hora de identificar a la persona de manera correcta. Es por eso que, el usar el umbral de 0.1 entrega un desempeño del 93 %. Por último, la métrica F1 simplifica las medidas de precisión y exhaustividad, donde va de 0 a 1, o 0 a 100 en porcentaje, siendo 1 (100) el mejor caso. El decir la misma frase tiene mejor rendimiento en el sistema, evaluándose con la medida F1 debido a que hay mayores verdaderos positivos a la hora evaluarse el sistema.

Analizando el sistema de identificación de voz dentro del marco de evaluación de la Tabla 2.3 al utilizar un umbral de 0.1, el usar la misma frase y una frase distinta, se obtiene un nivel medio. Esto debido a que, se requiere un valor entre 3 a 10 % en alguno de los valores de FRR o FAR para ser clasificado en ese nivel. Por lo tanto, no hay una relevancia significativa en utilizar una frase distinta y la misma de registro al momento de requerir precisión en el sistema o un bajo FAR, ya que ambos están por debajo de 10 %.

Por otro lado, al observar el sistema de identificación por voz, mediante la métrica de FRR, se observa que, hay una gran significancia entre utilizar la misma frase a una distinta, debido a que el FRR está relacionado con la exactitud del sistema. Es decir, mientras la exactitud va aumentando, el FRR va disminuyendo, teniendo una correlación inversa entre ambas métricas.

El FRR en el umbral 0.1 entre la misma frase y una distinta, tienen una diferencia de aproximadamente 26 %, indicando que el usar la misma frase tiene mayor probabilidad de asignar un valor positivo al momento de autentificarse. Esto sin considerar si este valor, es un falso positivo o no.

Capítulo 6

Conclusiones y trabajos a futuro

En los siguientes apartados de este capítulo, se describen tres puntos, las conclusiones a la que llega el autor del trabajo desarrollado, las aportaciones que se obtuvieron y que se espera realizar como trabajos a futuro. De esta manera, en los apartados siguientes se puede tener noción de que es lo más relevante del trabajo y desde el punto de vista del autor se puede tener nuevos caminos a seguir para continuar con esta línea de investigación.

6.1. Conclusiones

Durante el desarrollo del trabajo de tesis, se entendió la complejidad del análisis de la voz. Se estudio la importancia de los sistemas biométricos como métodos de autentificación, además que se comprendió el funcionamiento completo de cualquier tipo de sistema biométrico. La implementación de la identificación por voz tuvo como finalidad el explorar la confiabilidad de un sistema de reconocimiento por voz.

El análisis de umbrales que se implementó con las distintas metricas tanto las de clasificación como las utilizadas para sistemas biométricos, nos da un mejor panorama de que tan bueno es un sistema biometrico. Ahora bien, en el caso de la implementación desarrollada, el utilizar umbrales bajos con base en los resultados nos da la certeza que tiene una alta confiabilidad por el hecho de que entre más cercano a 0 sea el umbral más similar debe ser la voz para que sea autentificada.

Por otra parte, el uso de la misma frase de registro a una distinta para realizar la autentificación, se tiene menor probabilidad de llegar al umbral permitido, pero con la misma

confiabilidad de aproximadamente 93 % de ser la persona reconocida. Esto se debe a que repetir una misma frase tiene mayor exactitud de emitir las mismas frecuencias, recordando que es con lo que se trabajó para el reconocimiento.

El reconocimiento de voz, sigue siendo un reto abierto como método de autentificación, ya que existen distintas variables a considerar que pueden influir bastante. El ruido ambiental, las emociones, las enfermedades, la distancia al micrófono, son algunos ejemplos de los factores que hacen complejo a este tipo de rasgo biometrico.

6.2. Aportaciones

Durante el desarrollo de esta investigación, se implementó un sistema de identificación por voz utilizando redes convolucionales. Ahora bien, con la finalidad de evaluar el sistema, se obtuvieron dos importantes contribuciones. El desarrollo de un *dataset* con características que cumplieran los requisitos para la evaluación del sistema implementado y el análisis de distintos umbrales para determinar la confiabilidad del sistema a la hora de autentificar una persona. Las aportaciones serán descritas a continuación:

- Desarrollo de *dataset*:

  El *dataset* está compuesto con frases iguales a la del registro y frases distintas. Esto con la finalidad de evaluar el impacto que tiene utilizar cada una de ellas al momento de autentificar mediante la voz y encontrar la mejor precisión. De la misma manera, el *dataset* fue diseñado con un ambiente controlado con el único propósito de reducir el ruido externo.

  El *dataset* cuenta con 128 personas etiquetadas y cada una con 6 audios, dando un total de 768 audios. La variedad de los audios está distribuida de dos formas, una de 384 audios con una misma frase y la otra son 384 audios con frases distintas seleccionadas aleatoriamente por el mismo usuario al momento de registrarse. El desarrollo completo del *dataset* está descrito en el capítulo 4.

- Análisis de umbrales en el sistema de identificación por voz:

El análisis de distintos umbrales de aceptación tuvo la finalidad de aumentar la precisión del sistema. El objetivo fue obtener una precisión arriba de 90 %, logrando obtener una precisión 93.79 % con un FAR de 9.57 %. Y en frases distintas se logró una precisión de 92.31 % y un FAR de 5.26 %.

6.3. Trabajos a Futuro

Con la finalidad de mejorar el sistema de autentificación por voz, se espera como trabajo a futuro explorar otras arquitecturas de aprendizaje profundo para aumentar la exactitud del sistema. A su vez, mantener e inclusive incrementar la precisión. De la misma manera, existen extractores de características como los Coeficientes Cepstrales de Mel y el banco de filtros, los cuales no fueron explorados para la extracción de características de audios. Por último, combinar estos extractores con distintas redes profundas para encontrar el mejor resultados para tener una buena clasificación.

Una vez mejorado el sistema de identificación de voz, se propone agregar una nueva funcionalidad donde se detecten audios falsos. Es decir, detectar cuando se reproducen audios de voz en la entrada del micrófono para así robustecer y atacar este tipo de problemas comunes en los sistemas de reconocimiento de voz.

Referencias

Ammour, B., Bouden, T. & Boubchir, L. (2018). Face-Iris Multimodal Biometric System using Multi-resolution Log-Gabor Filter with Spectral Regression Kernel Discriminant Analysis. *IET Biometrics*, *7*. https://doi.org/10.1049/iet-bmt.2017.0251 (page 25)

Arya, K. & Bhadoria, R. (2019). *The Biometric Computing: Recognition and Registration*. Taylor & Francis Group. https://books.google.com.mx/books?id=pr7TxgEACAAJ. (Page 24)

Bae, H.-S., Lee, H.-J. & Lee, S.-G. (2016). Voice recognition based on adaptive MFCC and deep learning. *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, 1542-1546 (pages 42, 44).

Bala, A., Kumar, A. & Birla, N. (2010). Voice command recognition system based on MFCC and DTW. *International Journal of Engineering Science and Technology*, *2*(12), 7335-7342 (page 36).

Banxico. (2021). Informe anual sobre el ejercicio de las atribuciones conferidas por la Ley para la Transparencia y Ordenamiento de los Servicios Financieros [Web; accedido el 13-04-2022]. %5CURL%7Bhttps://www.banxico.org.mx/publicaciones-y-

prensa/informes-anuales-de-cumplimiento-de-la-ley-para-la/%5C%7B673FB877-292B-1DF2-2703-140F9168FDA9%5C%7D.pdf%7D. (Page 1)

Boles, A. & Rad, P. (2017). Voice biometrics: Deep learning-based voiceprint authentication system. *2017 12th System of Systems Engineering Conference (SoSE)*, 1-6 (pages 42, 44).

Chatfield, K., Simonyan, K., Vedaldi, A. & Zisserman, A. (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets. https://doi.org/10.48550/ARXIV.1405.3531. (Page 55)

Chollet, F. (2015). Keras [Web; accedido el 13-04-2022]. %5CURL%7Bhttps://keras.io/about/%7D. (Pages 37, 38)

Dai, J. & Zhou, J. (2010). Multifeature-based high-resolution palmprint recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(5), 945-957 (page 27).

Daouk, C., El-Esber, L., Kammoun, F. & Al Alaoui, M. (2002). Iris recognition. *IEEE ISSPIT*, (4), 558 (page 26).

Das, T. & Nahar, K. (2016). A voice identification system using hidden markov model. *Indian Journal of Science and Technology*, *9*(4), 1-6 (pages 41, 44).

de Martin-Roche, D., Sanchez-Avila, C. & Sanchez-Reillo, R. (2001). Iris recognition for biometric identification using dyadic wavelet transform zero-crossing. *Proceedings IEEE 35th Annual 2001 International Carnahan Conference on Security Technology (Cat. No. 01CH37186)*, 272-277 (pages 32, 33).

Dhakal, P., Damacharla, P., Javaid, A. & Devabhaktuni, V. (2019). A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface. *Mach. Learn. Knowl. Extr.*, *1*, 504-520 (pages 40, 44).

Dongare, A., Kharde, R., Kachare, A. D. et al. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(1), 189-194 (page 10).

Gayathri, M., Malathy, C. & Prabhakaran, M. (2019). A Review on Various Biometric Techniques, Its Features, Methods, Security Issues and Application Areas. *International Conference On Computational Vision and Bio Inspired Computing*, 931-941 (pages 1, 2, 25).

Hanifa, R. M., Isa, K. & Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, *90*, 107005 (page 39).

Hashiyada, M. (2011). DNA biometrics. https://doi.org/10.5772/18139. (Page 27)

Hendryli, J. & Herwindiati, D. E. (2020). Voice authentication model for one-time password using deep learning models. *Proceedings of the 2020 2nd international Conference on Big Data Engineering and Technology*, 35-39 (page 44).

Jahangir, R., Teh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., Akhtar, M. Z. & Ali, I. (2020). Text-independent speaker identification through feature fusion and deep neural network. *IEEE Access*, *8*, 32187-32202 (pages 40, 44).

Jain, A. K. & Li, S. Z. (2011). *Handbook of face recognition* (Vol. 1). Springer. (Page 26).

Jain, A. K. & Nandakumar, K. (2012). Biometric authentication: System security and user privacy. *Computer*, *45*(11), 87-92 (page 30).

Jangir, J., Singh, B. K. & Ali, M. I. (2014). Voice Identification Secure System by Statistical Model of Speech Signal Using Normalization Technique (pages 41, 44).

Joshi, M., Mazumdar, B. & Dey, S. (2018). Security Vulnerabilities Against Fingerprint Biometric System. https://doi.org/10.48550/ARXIV.1805.07116. (Page 25)

Kalimoldayev, M., Mamyrbayev, O. Z., Kydyrbekova, A. & Mekebayev, N. (2019). Voice verification and identification using i-vector representation. *International Journal of Mathematics and Physics*, *10*(1), 66-74 (pages 43, 44).

Kong, A., Zhang, D. & Kamel, M. (2009). A survey of palmprint recognition. *Pattern Recognition*, *42*(7), 1408-1418. https://doi.org/https://doi.org/10.1016/j.patcog.2009.01.018 (page 26)

Lukic, Y., Vogt, C., Dürr, O. & Stadelmann, T. (2016). Speaker identification and clustering using convolutional neural networks. *2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)*, 1-6 (pages 43, 44).

Marasco, E. & Ross, A. (2014). A survey on antispoofing schemes for fingerprint recognition systems. *ACM Computing Surveys (CSUR)*, *47*(2), 1-36 (page 25).

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, . . . Xiaoqiang Zheng. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Software available from tensorflow.org]. https://www.tensorflow.org/. (Page 38)

Minaee, S., Abdolrashidi, A., Su, H., Bennamoun, M. & Zhang, D. (2019). Biometrics recognition using deep learning: A survey. *arXiv preprint arXiv:1912.00271* (pages 2, 33).

Nagrani, A., Chung, J. S. & Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (pages 24, 54).

Negnevitsky, M. (2005). A guide to intelligent systems. *Artificial Intelligence, 2nd edition, pearson Education* (pages 7-13).

Nilu, S., Agrawal, A. & Khan, P. R. (2018). Voice Biometric: A Technology for Voice Based Authentication. *Advanced Science, Engineering and Medicine*, *10*. https://doi.org/10.1166/asem.2018.2219 (page 28)

Omil, J. C. (2019). Inteligencia artificial¿ Dr. Jekyll o Mr. Hyde? *Mercados y Negocios*, (40), 5-22 (page 7).

Paszke, A. (s.f.). PyTorch [Web; accedido el 13-04-2022]. %5CURL%7Bhttps://pytorch.org/docs/stable/index.html%7D. (Page 38)

Patterson, J. & Gibson, A. (2017). *Deep learning: A practitioner's approach*. .º'Reilly Media, Inc." (Pages 11, 14, 16-23).

Petrovska-Delacrétaz, D., Chollet, G. & Dorizzi, B. (2009). *Guide to biometric reference systems and performance evaluation*. Springer. (Page 32).

Plchot, O., Burget, L., Aronowitz, H. & Matejka, P. (2016). Audio enhancing with DNN autoencoder for speaker recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5090-5094 (pages 42, 44).

Querini, M., Gattelli, M., Gentile, V. & Italiano, G. F. (2015). A new system for secure handwritten signing of documents (page 29).

Rui, Z. & Yan, Z. (2018). A survey on biometric authentication: Toward secure and privacy-preserving identification. *IEEE access*, *7*, 5994-6009 (pages 31, 48).

Singh, J. P., Jain, S., Arora, S. & Singh, U. P. (2018). Vision-Based Gait Recognition: A Survey. *IEEE Access*, *6*, 70497-70527. https://doi.org/10.1109/ACCESS.2018.2879896 (page 28)

Tamoto, A. & Itou, K. (2019). Voice authentication by text dependent single utterance for in-car environment. *Proceedings of the Tenth International Symposium on Information and Communication Technology*, 336-341 (pages 41, 44).

Tandel, N. H., Prajapati, H. B. & Dabhi, V. K. (2020). Voice Recognition and Voice Comparison using Machine Learning Techniques: A Survey. *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 459-465 (pages 2, 33, 34).

Torfi, A., Dawson, J. & Nasrabadi, N. M. (2018). Text-independent speaker verification using 3d convolutional neural networks. *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 1-6 (pages 43, 44).

van Rossum, G. (1991). Python [Web; accedido el 13-04-2022]. %5CURL%7Bwww.python.org/about/%7D. (Page 37)

Variani, E., Lei, X., McDermott, E., Moreno, I. L. & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 4052-4056 (pages 42, 44).

Wang, S.-C. (2003). Artificial neural network. *Interdisciplinary computing in java programming* (pp. 81-100). Springer. (Page 9).

Zhang, Y., Huang, Y., Wang, L. & Yu, S. (2019). A comprehensive study on gait biometrics using a joint CNN-based method. *Pattern Recognition*, *93*, 228-236. https://doi.org/https://doi.org/10.1016/j.patcog.2019.04.023 (page 28)

Zhu, Y., Tan, T. & Wang, Y. (2000). Biometric personal identification based on handwriting. *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, *2*, 797-800 (page 29).

Zulfiqar, M., Syed, F., Khan, M. J. & Khurshid, K. (2019). Deep Face Recognition for Biometric Authentication. *2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 1-6. https://doi.org/10.1109/ICECCE47252.2019.8940725 (page 26)