SISTEMA DE RECONOCIMIENTO DE VOZ: UNA APLICACIÓN PARA AUMENTAR LA EFICIENCIA DEL SISTEMA ESPERANZA

MAGISTER EN MATAMÁTICAS APLICADAS Y CIENCIAS DE LA COMPUTACIÓN

WILLIAM LIZARAZO MALAMBO

DIRECTORA DE TESIS

YIBY KAROLINA MORALES PINTO

Universidad del Rosario

Escuela de Ingeniería, Ciencia y Tecnología.

Maestría en Matemáticas Aplicadas y Ciencias de la Computación.

2024

Resumen

La Ley 906 de 2004 otorga a los fiscales la facultad para interceptar comunicaciones, bajo el cumplimiento de los requisitos legales. Esta herramienta, clave en investigaciones penales, enfrenta desafíos crecientes debido a la adopción de las tecnologías de comunicación encriptada y el alto volumen de datos. En este contexto, surge la necesidad de implementar nuevas metodologías de análisis de datos que incrementen la eficiencia del sistema de interceptación de comunicaciones. El reconocimiento e identificación de voz es una de estas metodologías, permitiendo a través de la generación de una huella digital, identificar y rastrear a la persona a la que pertenece dicha huella dentro de un conjunto de datos. Esta técnica resulta particularmente útil en el contexto judicial, ya que facilita la asociación de casos y mejora la eficiencia del sistema al identificar si una misma persona está siendo o ha sido monitoreada en diferentes salas posibilitando la construcción de casos más sólidos. La base técnica de este método incluye el análisis de frecuencias de sonido y el uso de espectrogramas, que actúan como huellas digitales en la identificación de voces. Para el análisis de estos datos se emplean dos modelos de redes neuronales convolucionales, modelos que son ampliamente usados para el análisis de este tipo de datos no estructurados. Las métricas de desempeño calculadas para cada uno de los modelos y experimentos diseñados muestran resultados satisfactorios para la solución del problema de identificación del hablante, sin embargo, uno de los modelos planteados domina todas las métricas aplicadas, siendo este el candidato para su implementación.

Abstract

Law 906 of 2004 grants prosecutors the power to intercept communications, subject to compliance with legal requirements. This tool, key in criminal investigations, faces growing challenges due to the adoption of encrypted communication technologies and the high volume of data. In this context, the need arises to implement new data analysis methodologies that increase the efficiency of the communications interception system. Voice recognition and identification is one of these methodologies, allowing, through the generation of a digital fingerprint, to identify and track the person to whom said fingerprint belongs within a set of data. This technique is particularly useful in the judicial context, since it facilitates the association of cases and improves the efficiency of the system by identifying whether the same person is being or has been monitored in different rooms, enabling the construction of more solid cases. The technical basis of this method includes the analysis of sound frequencies and the use of spectrograms, which act as fingerprints in the identification of voices. To analyze this data, two convolutional neural network models are used, models that are widely used for the analysis of this type of unstructured data. The performance metrics calculated for each of the designed models and experiments show satisfactory results for

solving the speaker identification problem, however, one of the proposed models dominates all the applied metrics, making this the candidate for implementation.

Índice

Índice de cuadros

Lista de figuras

1. JUSTIFICACIÓN Y DESCRIPCIÓN DEL PROBLEMA

La administración de justicia en Colombia, y potencialmente en muchos otros países, enfrenta un desafío significativo: la cantidad de información a procesar supera ampliamente la capacidad de procesamiento de las entidades judiciales. Como resultado, el déficit en el procesamiento de información crece exponencialmente cada día, empujando al sistema judicial hacia un colapso inminente. Para abordar este problema, se han desarrollado varios programas de política pública.

Algunos de estos programas, orientados a la demanda, buscan proponer métodos alternativos de resolución de conflictos que eviten sobrecargar el sistema judicial, o mejorar la calidad de vida para reducir la cantidad de conflictos entre los ciudadanos. En contraste, los esfuerzos en el lado de la oferta se concentran en tres frentes. El primero se enfoca en cambios legislativos necesarios para agilizar la justicia, como la transición de procesos escritos a procesos orales. El segundo busca la adquisición y asignación de recursos, incluyendo el aumento de personal. Finalmente, el tercer frente se relaciona con la tecnología, buscando implementar nuevas soluciones que ayuden a gestionar y procesar la ingente cantidad de información que ingresa diariamente al sistema judicial.

A diferencia de la estrategia de incremento de personal, las acciones relacionadas con la adopción de nuevas tecnologías han demostrado un impacto significativo en la reducción de los tiempos procesales. Por esta razón, este documento propone el desarrollo y adopción de una tecnología de reconocimiento de voz para las salas de interceptación de la Fiscaliía General de la Nación (FGN). Esta tecnología, a través del registro de una huella digital de voz, permite establecer si un individuo investigado ha sido o está siendo monitoreado por algún componente del sistema de interceptación de comunicaciones de la FGN. Este proceso de identificación puede también asistir en la asociación de casos, ya que la huella de voz puede revelar conexiones entre distintos casos donde se haya detectado la misma voz. Además, la adopción de esta tecnología no solo permitirá mejorar la eficiencia del sistema de interceptación, sino que también facilitará la identificación de pruebas potencialmente valiosas que permitan a los fiscales formular acusaciones más sólidas.

Para contextualizar el problema específico al que se dirige esta tesis, es necesario describir en términos amplios el sistema de interceptación de la FGN. Cabe señalar inicialmente que “el derecho fundamental a la intimidad, consagrado en el artículo 15 de la Constitución Política, es una garantía inherente a un Estado social de derecho. Sin embargo, como la Corte Constitucional ha señalado, no se trata de un derecho absoluto, ya que puede ser limitado en circunstancias particulares por razones de interés

general, legítimas y debidamente justificadas constitucionalmente"[2].

Además, en la directiva 004 de 2022 [3] se definen los pasos legales, técnicos y procedimentales para solicitar una interceptación de comunicaciones. Siguiendo la descripción proporcionada por la revista Semana[4], el sistema de interceptación de la FGN consta de casi 30 salas que se dividen según la especialidad criminal y una distribución geográfica que abarca las regiones con los niveles más altos de conflictividad. Un coordinador general supervisa el sistema, verifica los procesos, da seguimiento a los mismos y asigna los recursos investigativos solicitados mediante una orden a la policía judicial, emitida por el fiscal encargado del caso.

Esta problemática se intensifica aún más debido a la arquitectura actual del sistema de interceptación. Los coordinadores de las salas solo tienen acceso a los números de los abonados telefónicos interceptados y al contexto general de la investigación. Sin embargo, no tienen acceso a la información detallada capturada y analizada por cada uno de los analistas, ya que esta es de uso exclusivo del fiscal que solicitó la interceptación de la o las líneas investigadas mediante una orden judicial.

Esta situación desemboca en un sistema de justicia que opera en silos de información, donde la información crucial para la resolución de los casos se mantiene aislada, impidiendo una visión holística de la situación y la posibilidad de realizar conexiones potencialmente significativas entre los casos. Esta estructura subyacente necesita ser abordada para permitir un flujo más eficiente de información y una mayor colaboración entre los fiscales, lo que a su vez podría conducir a casos más fuertes y una administración de justicia más eficaz y célere.

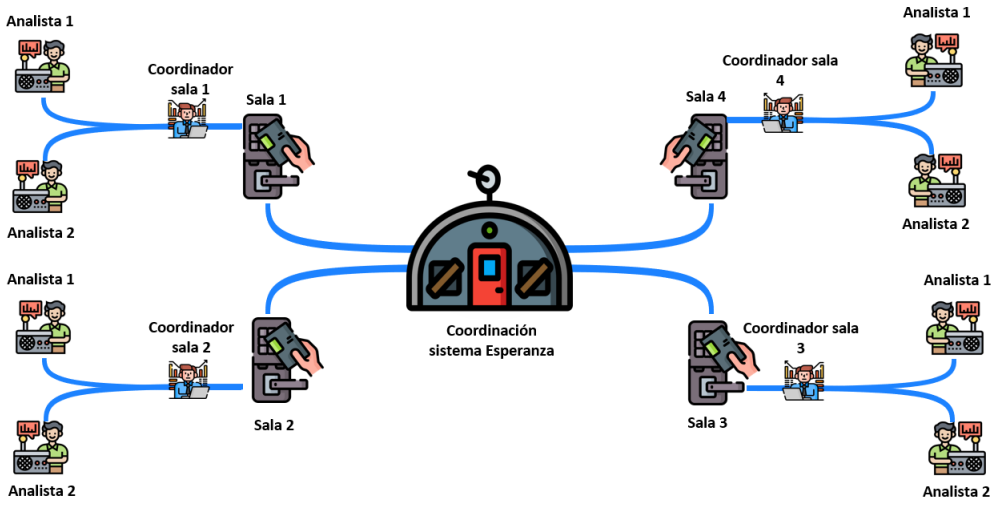
La Figura 1 proporciona un esquema general de la organización del sistema de interceptación. Como se mencionó, el sistema se compone de una serie de salas especializadas en distintas áreas de la investigación criminal, tales como narcotráfico, secuestro, extorsión, delitos violentos, antiterrorismo, entre otros. Cada una de estas salas cuenta con varios analistas, quienes se encargan de múltiples casos y cuya tarea principal es analizar las comunicaciones interceptadas legalmente e informar sus hallazgos al fiscal y/o al investigador principal del caso. Salvo en situaciones donde la vida de un individuo esté en riesgo inminente, o cuando existe una orden de inspección al proceso emitida por otro fiscal, el analista no tiene permitido compartir información con el coordinador de la sala, el coordinador del sistema, ni con ninguna otra persona que no esté directamente involucrada en la investigación del caso.

El coordinador de cada sala, entre otras funciones, debe supervisar el cumplimiento de los protocolos

establecidos para la interceptación de comunicaciones, y llevar a cabo tareas administrativas relacionadas con el tiempo y la evaluación de los informes de los analistas. Por su parte, el coordinador del sistema de interceptación tiene la responsabilidad de garantizar que el sistema funcione correctamente, verificar el cumplimiento de los protocolos y gestionar otras tareas administrativas pertinentes.

Esta descripción subraya la rigurosidad de las medidas de seguridad implementadas para la información capturada por el sistema y las barreras existentes para compartir dicha información.

Figura 1: arquitectura del sistema Esperanza



Fuente: elaboración propia basado en [4]

2. OBJETIVOS

En esta sección, se propone delinear de manera detallada tanto el objetivo principal como los objetivos específicos que guían esta investigación. Cabe subrayar la importancia y relevancia de este estudio: proponer una solución a un problema concreto que se ha identificado en el ámbito práctico de la investigación criminal. La originalidad y fortaleza de este enfoque radica en la implementación de metodologías destinadas al análisis de datos de audio. En particular, nos concentramos en sonido de voces como eje central.

2.1. Objetivo general

El objetivo de este trabajo radica en la formulación de una metodología y herramienta de reconocimiento de voz específicamente diseñada para el contexto de las salas de interceptación de la FGN. Esta herramienta tiene como objetivo principal la detección e identificación de nuevas señales de voz a partir de una base de datos preexistente.

Para materializar este objetivo, la herramienta propuesta se fundamentará en el uso de algoritmos de redes neuronales convolucionales. Estas redes son reconocidas por su habilidad para detectar patrones complejos como los que se hallan en las señales de voz. Para abordar el problema y lograr el objetivo, el uso de estas herramientas se realizará con base en su capacidad para abordar un problema de identificación del hablante. Con este enfoque, se podrá determinar si un individuo ha sido o está siendo monitoreado en más de una sala, revelando así potenciales conexiones entre casos que anteriormente permanecían ocultas.

La creación de esta aplicación busca no solo mejorar la eficiencia del sistema de interceptación, sino también fortalecer la calidad de las investigaciones al proporcionar un medio para asociar casos que involucran a un mismo individuo. Esta iniciativa se traducirá en un sistema judicial más efectivo, capaz de consolidar pruebas más robustas y de manejar la información de manera más integrada.

La implementación detallada de esta tecnología se examinará en profundidad en las siguientes cinco secciones. La primera sección, ya presentada, establece el contexto y la magnitud del problema que enfrentamos. La segunda sección articula el objetivo principal de esta tesis, delineando cada paso crítico necesario para alcanzar tal fin, en la tercera sección, dedicada al marco teórico, se lleva a cabo una revisión de la literatura que permite trazar la evolución de los sistemas de reconocimiento de voz, además de desarrollar la base teórica sobre la cual se construye la herramienta que resolverá el

problema planteado.

La cuarta sección se centra en el desarrollo de la herramienta, donde se describen tanto el conjunto de datos de prueba que se utilizarán como la arquitectura de los algoritmos que constituye la herramienta. A partir de los resultados obtenidos con la aplicación del modelo, se calculan diversas métricas de desempeño. Finalmente, en la última sección se discuten las conclusiones, así como las posibles mejoras y ajustes para futuros trabajos.

2.2. Objetivos específicos

Con el propósito de alcanzar el objetivo general, a continuación, se enumeran y edtallan los objetivos específicos. Estos servirán como hoja de ruta, delineando cada etapa y componente para la construcción de la solución propuesta.

1. Implementar y ajustar, al menos, dos modelos de predicción para el análisis de señales de sonido utilizando redes neuronales convolucionales.

2. Plantear una arquitectura de red específica para cada modelo propuesto que garantice un proceso de predicción de alta precisión y calidad.

3. Evaluar el desempeño de predicción de cada modelo propuesto y seleccionar aquel que ofrezca el mejor rendimiento para abordar el problema planteado.

4. Proponer un sistema que permita su implementación y uso dentro del sistema Esperanza

z

3. MARCO TEÓRICO Y ESTADO DEL ARTE

Este estudio surge de la curiosidad por entender el mecanismo o algoritmo fundamental que permite el reconocimiento e identificación de música que proporciona la tecnología de la aplicación Shazam. En 2003, Avery Li-Chun Wang patentó un algoritmo que se caracteriza por ser resistente al ruido y a la distorsión, computacionalmente eficiente y altamente escalable, capaz de identificar rápidamente un fragmento corto de música capturado a través de un micrófono de teléfono móvil en presencia de voces en primer plano y otros ruidos dominantes, incluso cuando se transmite a través de códecs de compresión de voz, a partir de una base de datos de más de un millón de pistas. El algoritmo emplea un análisis de constelación tiempo-frecuencia con hash combinatorio del audio, lo que da lugar a propiedades inusuales como la transparencia, en la que se pueden identificar múltiples pistas mezcladas. Además, para aplicaciones como la monitorización de radio, los tiempos de búsqueda son del orden de unos pocos milisegundos por consulta, incluso en una base de datos musical masiva"[5]. Este desarrollo seminal sentó las bases para la generación de nuevas metodologías que mejoran la eficiencia de la búsqueda y su precisión. No obstante, todas estas mejoras se asientan en un principio fundamental: la clasificación de la información a partir del análisis de espectrogramas digitales, también conocidos como "huellas de voz".

Este trabajo se ubica dentro del amplio espectro de literatura relacionada con el análisis de voz, un campo con múltiples aplicaciones como: el reconocimiento de discurso, que se enfoca en la traducción de las ondas sonoras producidas por un hablante a texto; traducción simultánea; verificación de hablantes; e identificación de hablantes. En este documento, el enfoque se centrará en las dos últimas aplicaciones, en el ámbito de la identificación y clasificación de sonidos.

Así mismo la verificación e identificación de hablantes tienen diferencias fundamentales, por ejemplo, la verificación de hablantes consiste en la validación de una señal en particular, es decir, determinar si la señal que se desea verificar coincide con la esperada. Este enfoque se usa mucho en la industria de seguridad biométrica. Por otro lado, la identificación de hablantes se basa en comparar la voz de la persona a identificar con un conjunto de voces, y determinar a quién pertenece.

Varios trabajos notables han abordado estos temas. Salehhaffari, H. (2018) [6], por ejemplo, empleó redes neuronales convolucionales para desarrollar una aplicación de verificación de voz, alcanzando una precisión del 77 %. Por su parte, Wang, M., Sirlapu, T., y otros (2018) [7] propusieron una solución de seguridad doméstica basada en el reconocimiento de voz, logrando una precisión del 84 % en menos de

60 ms en un dispositivo móvil utilizando una red neuronal convolucional.

Yanjie Jia et al. (2020) [8] presentan un enfoque que utiliza un análisis estadístico de espectrogramas de corto tiempo para identificar rasgos de pronunciación consistentes entre hablantes individuales, particularmente en el contexto de los acentos chinos. Para este propósito, emplean una base de datos que comprende a 100 hablantes. Posteriormente, realizan el reconocimiento de los hablantes a través de una red neuronal basada en un mapa autoorganizado de características, aplicando un método adaptativo de agrupamiento. Esta técnica demuestra la capacidad de mejorar la eficiencia del entrenamiento, aumentar significativamente la velocidad de reconocimiento y mejorar la precisión en comparación con un conjunto de modelos propuestos en el mismo estudio.

Hourri, S., Nikolo, N. y Kharroubi, J. (2021) [9] propusieron un complemento para los algoritmos de reconocimiento e identificación de voz basados en Redes Neuronales Convolucionales (RNC), introduciendo un análisis de vectores de características que mejora significativamente los tiempos de búsqueda.

Por otro lado, Arshad, S., Haider, S., y Mughal, A. (2022) [10] presentaron un modelo de aprendizaje no supervisado para identificar, dentro de un conjunto de datos, a quién pertenece una voz determinada. Este enfoque tiene la ventaja de que no requiere una etiquetación previa del conjunto de datos.

MaCleod, C. (2022) [1] proporcionó un tutorial detallado que describe el método empleado por Shazam para la identificación y clasificación de canciones. El procedimiento implica la transformación de Fourier para calcular espectrogramas, seguida de la extracción de características de estos espectrogramas. A través de este proceso, Shazam logra clasificar correctamente una amplia gama de canciones en cuestión de segundos, manteniendo un alto grado de precisión.

Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., ... & Metze, F. (2017) [11] presentaron Deep Speaker, un sistema basado en redes neuronales que incorpora arquitecturas como Residual Convolutional Neural Network (ResCNN) y Gated Recurrent Unit (GRU). Esta red se aplica directamente a los espectrogramas derivados de la transformación de Fourier realizada en las señales vocales de cada individuo. Gracias a la notable capacidad de las redes neuronales convolucionales para el procesamiento y análisis de imágenes, logran extraer con precisión las singularidades presentes en cada espectrograma.

Dentro de estos espectrogramas, identifican varios atributos clave que permiten diferenciar a los ha-

blantes. Por ejemplo, prestan especial atención a los tonos fundamentales, que se manifiestan en las frecuencias más extremas y suelen ser particulares de cada persona. Además, observan las agrupaciones de tonos que evidencian las peculiaridades de un hablante al articular ciertas palabras o construcciones lingüísticas.

Para su entrenamiento utilizaron una técnica conocida como 'mean polling'. Este proceso implica aplicar una función de pérdida que mide la distancia entre puntos similares, permitiendo que estos avancen a la siguiente etapa de entrenamiento mientras que los puntos distantes son descartados. Este enfoque resultó ser altamente efectivo, ya que Deep Speaker logró superar de manera significativa a un modelo de línea de base i-vector impulsado por Deep Neural Networks (DNN). Es de resaltar la adaptabilidad interlingüística de Deep Speaker, ya que este sistema mostró una mejora significativa en el reconocimiento de hablantes de inglés, incluso cuando inicialmente fue entrenado en mandarín. Esto sugiere un alto grado de versatilidad y eficacia en su capacidad para identificar y verificar hablantes a través de diferentes idiomas.

Estas características se traducen en huellas sonoras digitales, inconfundibles para cada individuo. Son procesadas por la red neuronal, meticulosamente adaptada para este propósito. Después de identificar y procesar estas peculiaridades se utilizan estas características para clasificar e identificar con precisión a los hablantes vinculados a cada muestra vocal.

3.1. Redes neuronales convolucionales para el procesamiento de datos de señales de voz

Las redes neuronales convolucionales (CNN) se han establecido como una poderosa herramienta para tareas de computación visual, lo que permite avances en campos como el reconocimiento de imágenes, el análisis de video e incluso el diagnóstico médico. La principal característica de las CNN radica en su estructura única y modelo computacional, que está diseñado para aprender de forma automática y adaptativa jerarquías espaciales de características a partir de los datos visuales.

A diferencia de los métodos tradicionales de aprendizaje automático, las CNN pueden procesar datos de píxeles sin procesar y, a través de una serie de capas convolucionales, no lineales y de agrupación, aprender a reconocer patrones característicos en los datos ingresados.

Las características clave de las CNN, como el aprendizaje de características locales y la invariancia posicional, las hacen excepcionalmente buenas para manejar datos de imágenes. Son capaces de detectar

características visuales locales como bordes, texturas y colores, y ensamblarlos en formas y objetos más complejos. Su capacidad para reconocer estas características, independientemente de su posición en la imagen, es crucial para tareas como la detección de objetos y el reconocimiento de imágenes.

Además, las CNN son computacionalmente eficientes con una menor cantidad de parámetros en comparación con las redes totalmente conectadas, gracias a los pesos compartidos y las capas de agrupación. Esto reduce el riesgo de sobreajuste y mejora la escalabilidad cuando se trata de imágenes de alta resolución. Finalmente, la disponibilidad de numerosos modelos preentrenados, como VGGNet, ResNet o Inception permiten un rápido desarrollo e implementación de modelos de alta precisión

Es posible que surjan dudas sobre cómo opera exactamente una red neuronal convolucional y por qué, en este documento, se propone un método que parece necesitar el análisis de datos de imágenes cuando lo que buscamos es clasificar sonidos. ¿Qué razón hay detrás de esta inusual transición entre dos tipos de información tan dispares?

La inquietud es válida. A grandes rasgos, una red neuronal convolucional (RNC) es una variante de red neuronal artificial especializada en procesar datos que siguen una cuadrícula topológica. Esto significa que están diseñadas para manejar datos estructurados como matrices 2D, siendo perfectas para el análisis de imágenes. No obstante, y como se destaca en este trabajo, las RNC no se limitan solo a imágenes; también pueden procesar y aprender de señales de onda sonora. Estas redes son capaces de extraer características únicas de señales en una dimensión 1D o en el dominio del tiempo.

A continuación, se presentan los dos formatos de datos que se utilizarán en este trabajo: en onda y en imagen. Estos serán analizados mediante dos estructuras de redes neuronales convolucionales: una especializada en procesar datos en 1D y otra en datos 2D.

Primero que todo comencemos a explicar de que se trata y de que está compuesto un modelo de redes neuronales convoluciones. Como se mencionó al inicio de este apartado las redes neuronales convolucionales (CNN, por sus siglas en inglés) son un tipo de red neuronal artificial diseñada específicamente para reconocer patrones en datos con estructura de rejilla, como imágenes. La convolución es una operación matemática que permite a la red identificar características específicas en regiones localizadas de la entrada.

Componentes principales de una CNN:

1. **Capas de entrada:** Aquí es donde la CNN recibe la información de la imagen. En el caso de las imágenes, la entrada suele ser una matriz tridimensional compuesta por la altura, la anchura y la profundidad (canales) de la imagen, como RGB.

2. **Capas convolucionales:** Estas son las capas clave de una CNN. Durante la convolución, se toma un filtro (una pequeña matriz) y se desliza sobre la imagen de entrada para producir un mapa de características. Este proceso permite a la CNN detectar características locales como bordes, texturas, etc. La idea es que la red aprenda los valores óptimos del filtro que permiten identificar las características relevantes para la tarea en cuestión.

3. **Función de activación:** Después de cada capa convolucional, se aplica una función de activación. La función ReLU (Rectified Linear Unit) es la más utilizada. Su propósito es introducir no linealidad en el modelo, permitiendo que la red aprenda patrones más complejos.

4. **Capas de agrupación:** Estas capas reducen la dimensionalidad de los mapas de características, manteniendo la información esencial. Hay diferentes tipos de pooling, siendo el "max pooling" uno de los más comunes. En el max pooling, se selecciona el valor máximo de un área determinada del mapa de características.

5. **Capas totalmente conectadas:** Después de varias capas de convolución y pooling, las características aprendidas se pasan a una o más capas totalmente conectadas para realizar la clasificación final. Estas capas son similares a las de las redes neuronales tradicionales.

6. **Capa de salida:** Esta es la última capa, que produce las predicciones finales del modelo. Si se trata de una tarea de clasificación, por ejemplo, esta capa podría usar una función de activación softmax para distribuir las probabilidades entre las clases.

Ahora bien, aunque se ha mostrado que una gran parte de las aplicaciones de las CNN se centra en datos 2D, especialmente en imágenes, el alcance y versatilidad de estas redes no se limitan a este tipo de datos. La adaptabilidad inherente de las CNN permite que su metodología de análisis pueda ser extendida a otras estructuras de datos. Por ejemplo, en el análisis de series temporales, las CNN pueden ser empleadas para detectar patrones o anomalías a lo largo del tiempo, aprovechando su capacidad para aprender características locales. Del mismo modo, en el contexto de los grafos, las CNN pueden ser adaptadas para identificar subestructuras o patrones que puedan ser indicativos de ciertas propiedades

o características del grafo.

Para comprender el funcionamiento de las CNN en relación con los datos empleados en el desarrollo de la herramienta propuesta en este documento, se inicia explicando su interacción con datos unidimensionales (1D). Posteriormente, se aborda de manera similar su tratamiento con datos bidimensionales (2D).

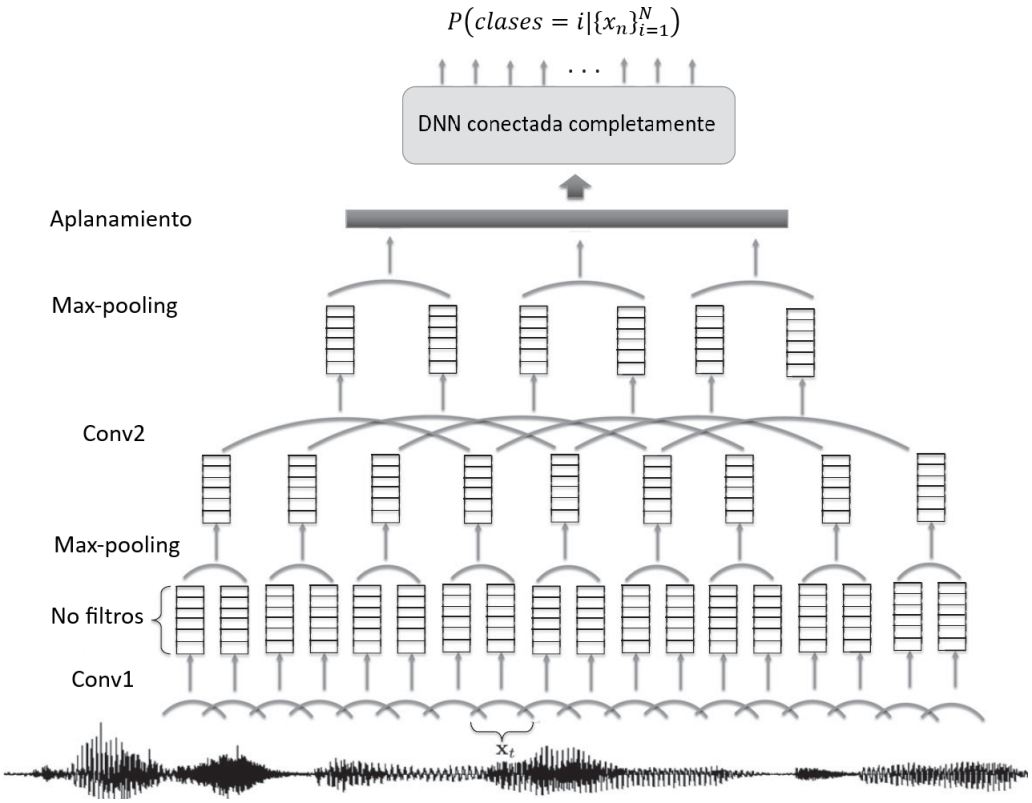
3.2. CNN para datos 1D

La voz, al igual que otras señales sonoras, posee características únicas que la diferencian de otros sonidos. En el ámbito del procesamiento de señales, y en particular en el análisis de voces, se han establecido procedimientos para resaltar estas distintivas propiedades. Una técnica destacada en este dominio es la de los coeficientes cepstrales de frecuencia mel (MFCC). Estos coeficientes representan el espectro de potencia de un sonido en un periodo corto y se han convertido en herramientas esenciales en áreas como el reconocimiento de voz y la recuperación de información musical. Un aspecto crucial al emplear los MFCC en el reconocimiento de voz es su capacidad para capturar la percepción no lineal de la frecuencia característica del oído humano. Esta es la razón por la cual discernimos con mayor facilidad entre frecuencias de 500 Hz y 1000 Hz que entre 10,000 Hz y 10,500 Hz. Precisamente, la escala Mel fue concebida para emular esta singular percepción frecuencial del oído humano.De igual manera, en la literatura especializada, se destacan otros enfoques relevantes en el procesamiento y análisis de señales de voz. Entre estos, la predicción lineal perceptual (PLP) y los códigos predictivos lineales emergen como métodos prominentes. La predicción lineal perceptual (PLP), inspirada en la percepción auditiva humana, busca modelar el habla aprovechando las características psicoacústicas del sistema auditivo [12]. Por otro lado, los códigos predictivos lineales se centran en representar la señal de habla en términos de una combinación lineal de sus muestras anteriores, ofreciendo una aproximación efectiva para la compresión y transmisión de voz. Ambos métodos, con sus respectivas ventajas, han demostrado ser cruciales en el campo del reconocimiento de voz.

Siguiendo esta línea, las señales de voz, cuando se analizan en el dominio del tiempo o de frecuencia, se presentan como señales unidimensionales (1D). La extracción de características en este dominio implica obtener atributos relevantes directamente de la serie temporal en bruto, sin necesidad de transformar previamente los datos a otro dominio, como podría ser el dominio de la frecuencia. En el caso del uso de las CNN para el análisis de este tipo de datos las capas convolucionales filtran los datos de series temporales por pequeños segmentos fijos sin procesar para detectar patrones temporales locales.

Siguiendo a Man-Wai Mak y Man-Wai Mak (2021) [5] en el siguiente gráfico se representa una arquitectura de una CNN para el procesamiento de una señal de sonido en dominio de tiempo.

Figura 2: CNN para el procesamiento de una señal de sonido



Fuente: Man-Wai Mak y Man-Wai Mak (2021) [13]

La anterior arquitectura puede ser descrita de la siguiente manera:

1. **Capa de entrada:** La capa de entrada es un vector donde cada valor represente la amplitud de la señal de audio en un momento dado. Por ejemplo, si utilizamos una frecuencia de muestreo de 16.000 Hz y tomamos una muestra de 1 segundo, la entrada sería un vector de longitud 16.000

2. **Capa convolucional 1:** Filtros: 18 filtros, Tamaño del filtro: 9 (esto significa que el filtro abarca 9 pasos de tiempo), paso o stride: 1 (El filtro se mueve paso a paso), Función de activación: ReLU (Unidad lineal rectificada). Esta capa escaneará la entrada en busca de patrones locales utilizando 18 filtros diferentes.

3. **Capa pooling:** En esta capa se aplica una función del tipo Max-pooling la cual maximiza las

caracteristicas más relevantes. El usuario define el tamaño del pool. por ejejmplo, esta capa reducirá la muestra de los mapas de características tomando el valor máximo cada 4 pasos de tiempo, comprimiendo así los datos.

4. **Capa convolucional 2:** En la segunda capa convolucional se filtra la salida de la capa de agrupación anterior, identificando patrones más abstractos. En esta capa se especifica el numero y tamaño de filtros, los pasos o stride y la función de activación, por ejemplo ReLU.

5. **Capa pooling:** Nuevamente se aplica el filtro Max-pooling a los datos de salida de la anterior capa, manteniendo el tamaño del pool.

6. **Capa de aplanamiento:** Esta capa transformará la salida 2D de la capa de agrupación anterior en un vector 1D, lo que la hará adecuada para la entrada en una capa completamente conectada. Por ejemplo, después de pasar una imagen o datos de una serie temporal a través de una serie de capas convolucionales y de agrupación en una CNN, la salida suele ser un tensor multidimensional (por ejemplo, 2D para imágenes, posiblemente 3D si se consideran múltiples mapas de características o canales). Antes de que estos datos puedan introducirse en una capa de red neuronal completamente conectada, es necesario remodelarlos en un vector 1D. La capa de aplanamiento se encarga de esta tarea.

7. **Capa conectada completamente:** Esta es una capa de red neuronal estándar que procesará las características extraídas por las capas convolucionales. Esta compuesta por un número de neuronas y una función de activación que puede ser ReLu.

8. **Capa de salida:** Dado que el problema es de clasificación, esta capa está compuesta por n neuronas que representan el número de clases y, una función de activación que puede ser softmax.

Como se detallará en la siguiente sección, cuando se utilizan datos unidimensionales (1D) en combinación con un modelo de red neuronal convolucional (CNN), se implementa un preprocesamiento de la señal en estudio. Este paso previo tiene como objetivo potenciar las características de la señal procesada y, en consecuencia, mejorar la precisión del modelo. Un método comúnmente utilizado para este fin es la Transformada Rápida de Fourier. Esta transformación nos permite examinar la señal de audio en el dominio de la frecuencia, un espacio donde ciertas características clave pueden ser más fácilmente detectadas en comparación con el dominio del tiempo. En general, una señal de audio suele constar de

múltiples frecuencias entrelazadas. Al pasar al dominio de la frecuencia, es posible descomponer estos componentes individuales, simplificando así el análisis o la manipulación de la señal. Este enfoque es especialmente útil cuando se quiere realzar características particulares en una señal.

3.3. CNN para datos 2D

Cuando se trata de datos bidimensionales (2D), el flujo de procesamiento experimenta algunas diferencias significativas, especialmente en las etapas de ingesta y preprocesamiento de datos. La aplicación de una red neuronal convolucional (CNN) en este contexto se justifica principalmente por las características del espectrograma. Un espectrograma es una representación gráfica que muestra las distintas frecuencias que componen una señal de audio a lo largo del tiempo. Esta transformación convierte el audio, originalmente un conjunto de datos unidimensionales (1D), en una 'imagen', que es un conjunto de datos bidimensionales (2D).

Así, aunque el objetivo final es clasificar sonidos, se necesita analizar imágenes porque los datos de sonido se representan de manera más efectiva en el espacio 2D de un espectrograma. Esta transformación de sonido a imagen permite el uso de técnicas de visión por computadora y aprendizaje profundo, como las CNN, para analizar y clasificar las señales de sonido.

Antes de profundizar en cómo las Redes Neuronales Convolucionales (RNC) funcionan en la clasificación de sonidos, es esencial comprender cómo se transforma una señal de sonido, que inicialmente es una representación unidimensional, en una imagen, o una representación bidimensional. Este proceso es fundamental para el análisis que se realizará y para la metodología que se está proponiendo.

Para entender mejor el proceso de análisis que se aplica en este documento, es necesario introducir dos elementos clave: la Transformada de Fourier previamente discutida, y el espectrograma. Estos elementos nos permiten visualizar una señal desde dos perspectivas diferentes: el dominio del tiempo y el dominio de la frecuencia. En el primer caso, el dominio del tiempo, una señal se define principalmente por cómo varía a lo largo del tiempo. Tomando como ejemplo una señal de audio, esta se describe inicialmente en función de las fluctuaciones en la presión del aire, es decir, la onda sonora, conforme pasa el tiempo. En esta representación, el eje X simboliza el tiempo y el eje Y refleja la amplitud de la señal en cada momento específico.

La Transformada de Fourier entra en juego como un mecanismo de transición entre estos dos dominios, habilitando el análisis de la señal desde una perspectiva diferente: su espectro de frecuencia. En el

dominio de la frecuencia describe la misma señal en términos de las frecuencias que la componen. Cuando aplicamos la Transformada de Fourier a una señal en el dominio del tiempo, obtenemos su representación en el dominio de la frecuencia. En este último, el eje X denota las frecuencias (es decir, cuántas veces se repite una función periódica por unidad de tiempo) y el eje Y indica la amplitud de cada componente de frecuencia, o en otras palabras, cuánta presencia tiene cada frecuencia en la señal.

En el contexto del audio, si un componente de frecuencia (por ejemplo, 440 Hz, que corresponde a una nota alta) es fuerte (tiene una gran amplitud en el eje Y en el dominio de la frecuencia), significa que esta frecuencia es un componente predominante del sonido y será más audible al reproducirlo.

Para ilustrar, las dos siguientes gráficas muestran diferentes tipos de ondas de sonido. La primera gráfica es una onda sinusoidal con una frecuencia relativamente amplia, mientras que la segunda gráfica representa una onda digital que corresponde a una canción. Como se puede observar, esta última señal es más compleja que la primera, lo que hace que su clasificación usando métodos tradicionales de emparejamiento de ondas sea menos efectiva. En este punto, la Transformada de Fourier y el concepto del espectrograma cobran importancia, ya que permiten descomponer y visualizar señales complejas, facilitando su posterior análisis y clasificación.

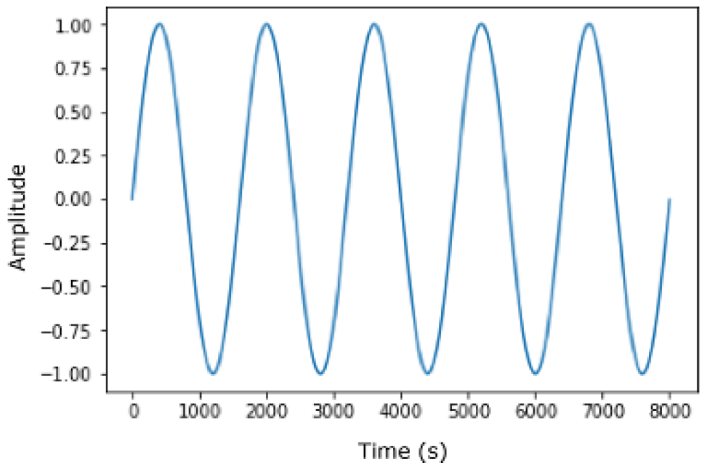


Figura 3: Representación de una onda sinusoidal
Fuente: Macleoad, Cameron (2022) [1]

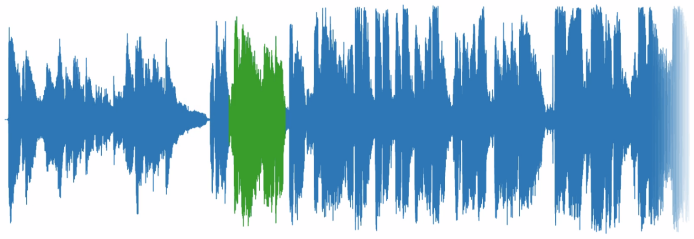


Figura 4: Representación de una onda digital de una canción
Fuente: Macleoad, Cameron (2022) [1]

El uso de la representación en el dominio de la frecuencia ofrece ciertas ventajas sobre la representación en el dominio del tiempo. Una de estas ventajas es que facilita la captura de las características esenciales de la señal analizada, permitiendo descartar el ruido que la acompaña. En contraste, aislar el ruido en el dominio del tiempo suele ser mucho más complicado y exigente desde el punto de vista computacional. Además, cuando una señal incorpora múltiples frecuencias, la identificación de estas se simplifica considerablemente en el dominio de la frecuencia comparado con el dominio del tiempo.

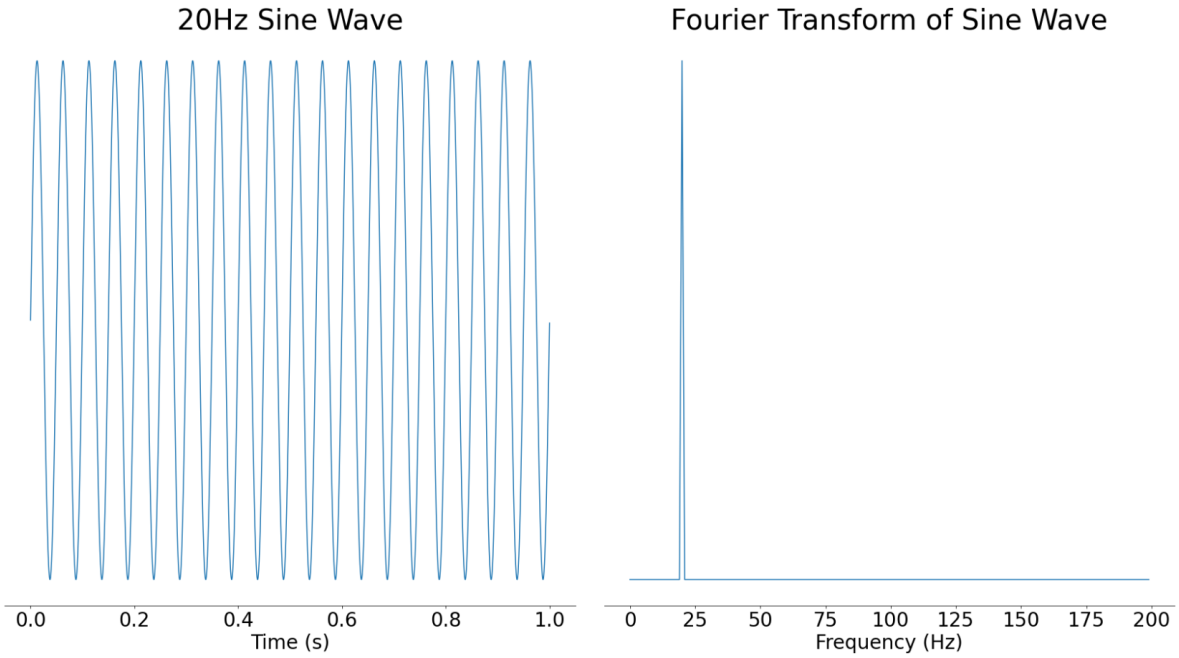


Figura 5: Resultado de aplicar la transformada de Fourier a una onda sinusoidal.
Fuente: Macleoad, Cameron (2022) [1]

Por las razones expuestas, las metodologías actuales para el análisis de sonido se centran en el estudio de los espectrogramas. Al aplicar una Transformada de Fourier a un fragmento de sonido, se pueden descifrar todas las diversas frecuencias que se manifiestan desde el principio hasta el final de la muestra,

proporcionando así una visión integrada de su contenido frecuencial. Sin embargo, este enfoque no facilita el reconocimiento de cómo la presencia o la intensidad de estas frecuencias fluctúan a lo largo de la pista de la muestra de sonido; por ejemplo, la introducción puede contener frecuencias dominantes distintas en ciertos fragmentos de la muestra análizada.

En este escenario es donde un espectrograma muestra su utilidad. Un espectrograma es una representación visual de las variaciones del espectro de frecuencias en un sonido u otra señal a lo largo del tiempo. Siguiendo a MacLeoad, Cameron (2022) [1] se presenta a continuación una serie de pasos para el cálculo y representación de un espectrograma. Este procedimiento inicia aplicando la Transformada de Fourier a pequeños segmentos o "ventanas"de una señal de manera sucesiva, proceso conocido como Transformada Rapida de Fourier de Tiempo Corto (STFFT, por sus siglas en ingles).

Este procedimiento genera una secuencia de espectros de frecuencia, cada uno correspondiente a una ventana de tiempo diferente de la señal analizada. Cada espectro proporciona información sobre qué frecuencias estaban presentes y cuán intensas eran durante ese intervalo de tiempo específico.

Al ensamblar todos estos espectros, se obtiene una gráfica tridimensional en la que el eje "xrepresenta el tiempo (desde el comienzo hasta el final de la señal), el eje zrepresenta la frecuencia (de menor a mayor), y el color o la intensidad en cada punto indican la fuerza o amplitud de una frecuencia en particular en un instante dado.

De esta forma, un espectrograma permite visualizar no sólo qué frecuencias están presentes en una señal de sonido, sino también cómo varía la presencia e intensidad de esas frecuencias a lo largo de una señal.

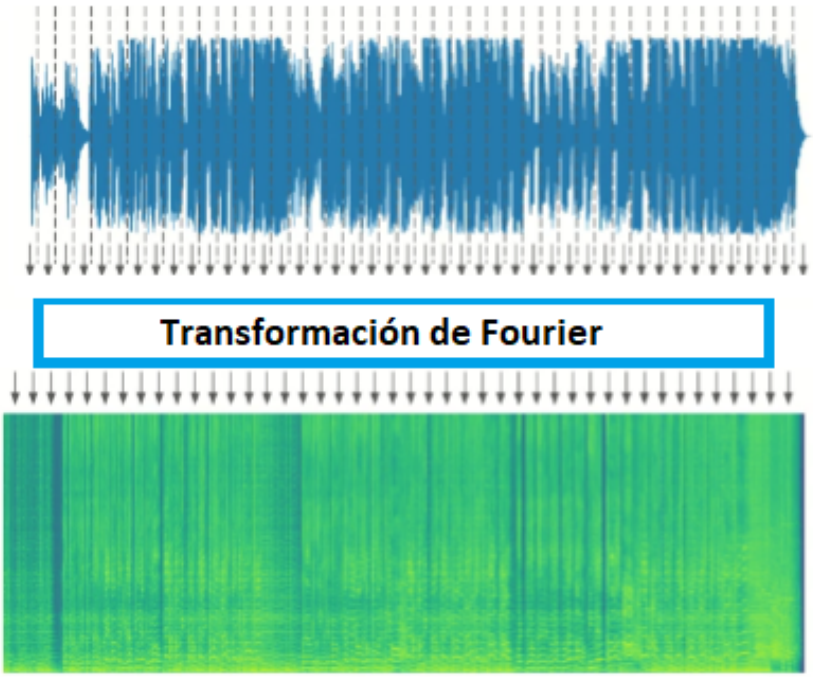


Figura 6: Proceso del cálculo del espectrograma de una canción.
Fuente: Macleoad, Cameron (2022) [1]

Tras generar el espectrograma de la señal objetivo, que es esencialmente una representación gráfica bidimensional de la señal de audio, podemos emplear métodos de clasificación de imágenes comúnmente utilizados, como las redes neuronales convolucionales. En este contexto, una CNN multicapa se encarga de procesar dicha información, extrayendo las características únicas que distinguen cada sonido. Estas características se identifican y organizan en diferentes capas de la red en lo que se conoce como "mapas de características". Posteriormente, estos mapas alimentan un clasificador lineal que se entrena mediante el ajuste de los pesos asignados a cada neurona. Este proceso iterativo continúa hasta alcanzar métricas de clasificación que satisfagan los criterios de aceptación preestablecidos, asegurando así la mayor precisión posible en la identificación del sonido.

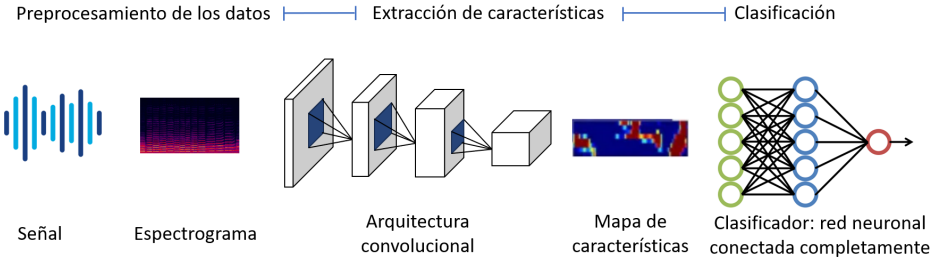


Figura 7: Proceso general de análisis y clasificación de sonidos por medio de una CNN
Elaboración propia basado en Doshi, Keatan (2021)[?]

Tal como se ha señalado, la estructura fundamental de una red neuronal convolucional (CNN) se mantiene en gran medida constante, independientemente de si se aplica al procesamiento de datos en una, dos o tres dimensiones (1D, 2D o 3D). La variación más notable se encuentra en la configuración de la capa de entrada. En el caso de datos bidimensionales, la entrada suele ser la imagen del espectrograma. Por ejemplo, si el espectrograma se ha generado con una resolución de 128×128 píxeles, esas serían las dimensiones de la capa de entrada de la CNN. A partir de esta capa inicial, las subsiguientes capas de la CNN mantienen una arquitectura idéntica, independientemente de la dimensionalidad de los datos.

4. Propuesta metodológica: arquitectura del sistema de análisis

En la siguiente sección, se aborda de manera exhaustiva las metodologías y enfoques adoptados para el tratamiento y la clasificación de datos de voz, poniendo un especial énfasis en el caso específico que es objeto de esta tesis. En primer lugar, se presentará una descripción detallada del conjunto de datos elegido, explicando su origen, tipo y las características intrínsecas que lo hacen idóneo para aplicar y probar el sistema propuesto. Seguidamente, se delinea la estructura completa del experimento, considerando variables como el número de archivos de voz a analizar y la longitud temporal de cada grabación individual. En tercer lugar, se establecen las métricas de desempeño que servirán como criterios de comparación y evaluación de los resultados. Posteriormente, se avanza a la fase de entrenamiento y ajuste de los modelos de Redes Neuronales Convolucionales (CNN) propuestos, calculando las métricas de desempeño para evaluar su eficacia. Finalmente, se consolida y se presenta un resumen de los hallazgos, proporcionando una perspectiva comprensiva del rendimiento de los modelos y las implicaciones para futuras investigaciones.

Una vez expuestos los algoritmos bases que se utilizaran para el desarrollo de la solución al problema planteado, Se presenta a continuación la arquitectura que permitirá la implementación de la solución dentro del sistema de interceptación de la FGN.

Es importante mencionar que ambas arquitecturas de redes que serán utilizadas y evaluadas corresponden a los siguientes criterios de elección:

1. **Naturaleza de los datos**: Dependiendo de la dimensionalidad y tipo de datos, la arquitectura adecuada varía. Datos unidimensionales, como señales de audio, se alinean bien con arquitecturas Conv1D, ya que estas están diseñadas para aprender características a lo largo de una dimensión. Por otro lado, datos bidimensionales como imágenes o espectrogramas se beneficiarán de Conv2D, que puede aprender características espaciales.

2. **Desempeño**: La metodología de "Backward Stepwise"se utiliza para simplificar modelos complejos. Esto significa que se comenzó con un modelo con muchas capas y características, y se simplificó eliminando iterativamente las menos importantes, basándose en métricas de evaluación. Esto asegura que el modelo final sea lo más eficiente posible, manteniendo un rendimiento óptimo.

3. **Naturaleza del problema**: El problema de clasificación supervisada exige que el modelo aprenda a distinguir entre diferentes categorías o clases a partir de los datos de entrada. Las arquitecturas seleccionadas están diseñadas para extraer características importantes de los datos y usarlas para hacer predicciones precisas sobre la categoría a la que pertenecen.

4. **Simplicidad**: Es esencial elegir una arquitectura que sea suficientemente simple para ser interpretable y fácil de entrenar, pero lo suficientemente compleja para capturar las características necesarias de los datos. La simplicidad también conlleva beneficios en términos de tiempo de entrenamiento y recursos computacionales. Es una balance entre rendimiento y coste computacional.

4.1. Descripción de los datos

Para el desarrollo del ejercicio empírico, se recurrió a grabaciones de diálogos en formato MP3 obtenidas del portal en línea https://audio-lingua.eu/. Este recurso proporciona acceso a narraciones breves compartidas por individuos de diversas partes del mundo, abarcando un espectro amplio de temas. Los colaboradores en estas grabaciones comprenden un amplio espectro de género, edades y nacionalidades, lo que aporta una diversidad significativa en términos de acentos y estilos de habla. En la Tabla 1 se presenta un resumen de las características temporales de las grabaciones utilizadas, incluyendo la duración total, la duración promedio por grabación, la desviación estándar, y los valores máximo y mínimo de las duraciones de las grabaciones.

Para el entrenamiento de los modelos, se seleccionó un conjunto de 125 archivos de audio en español, correspondientes a 30 individuos distintos. De forma colectiva, estas grabaciones representan un total de 2 horas, 27 minutos y 40 segundos, con una duración promedio de 1 minuto y 11 segundos por archivo. En cuanto a la distribución de género, se incluyeron 16 individuos masculinos y 14 femeninos, contribuyendo con 63 y 58 grabaciones respectivamente.

Cuadro 1: Estadísticas de duración de las grabaciones

	Tiempo total	Tiempo promedio	Desviación estándar	Mínimo	Máximo
Registro	02:27:40	00:01:11	00:00:41	00:00:12	00:03:48

Cuadro 2: Distribución de grabaciones por género

Sexo	n	Tiempo total	Promedio	Desviación estándar	Mínimo	Máximo
Hombre	16	01:11:40	00:01:07.1875	00:00:43.430980	00:00:12	00:03:30
Mujer	14	01:16:00	00:01:16.0000	00:00:40.229005	00:00:16	00:03:48

4.2. Selección y Preparación de los Datos

En esta subsección se expone la metodología para la selección y preparación de los conjuntos de datos utilizados en cada uno de los experimentos realizados. Como se mencionó, el primer paso para el cálculo de los espectrogramas implica dividir cada archivo de audio en fragmentos de igual duración. De este modo, los 125 archivos de audio seleccionados para este estudio se dividieron en segmentos de un segundo de duración. En la Figura 8 se muestra la distribución del número de segmentos de audio resultantes para cada uno de los 30 individuos considerados en este estudio.

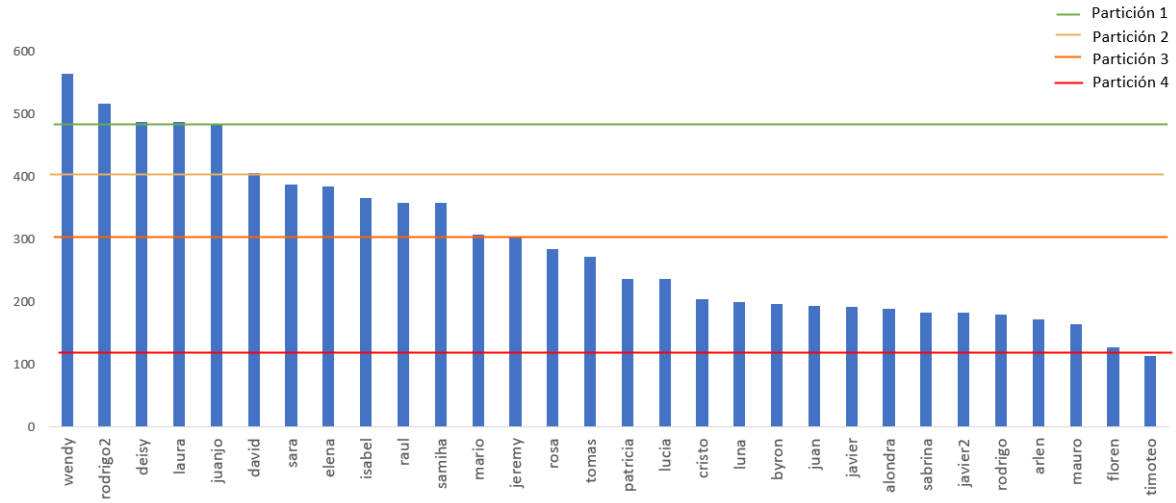


Figura 8: Número de archivos de audio por persona. (1 segundo/archivo)
Fuente: Cálculos propios

En el segundo paso para el entrenamiento del modelo se seleccionan individuos que posean una cantidad igual de archivos de audio. Este requisito se satisface mediante la selección de subgrupos de individuos, garantizando que todos en el subgrupo posean al menos un número específico de archivos de audio. Por ejemplo, nuestro primer subgrupo está conformado por individuos que disponen de al menos 483 archivos de audio de un segundo de duración, el segundo subgrupo contiene a aquellos con al menos 400 archivos, el tercer subgrupo agrupa a los que poseen un mínimo de 313 archivos, y finalmente, el cuarto subgrupo incorpora a todos los individuos que tienen, como mínimo, 113 archivos. Esta metodología permite estructurar subgrupos basados en la disponibilidad mínima de archivos de audio. De esta manera, se garantiza que cada subgrupo contenga una cantidad equiparable de datos para el entrenamiento del modelo. Esto contribuye a evitar la introducción de sesgos y asegura una representación equitativa y justa de cada individuo dentro del modelo. Además, la estrategia propuesta ofrece la oportunidad de someter al modelo a pruebas de estrés. Al reducir la cantidad de archivos disponibles, es decir, al trabajar con espectrogramas más condensados, se puede evaluar cómo fluctúan los indica-

dores de evaluación entre los diferentes subgrupos. Esta exploración proporciona valiosa información para entender la robustez y la tolerancia del modelo bajo condiciones de información limitada.

4.3. Arquitecturas de Redes Neuronales Convolucionales (CNN)

En esta subsección, se detallarán las arquitecturas específicas de los modelos de CNN empleados para el procesamiento y clasificación de los datos que constituyen el enfoque central de esta tesis. Se describirán los componentes clave de cada arquitectura, incluyendo las capas convolucionales, las capas de agrupamiento (pooling), y las capas totalmente conectadas, entre otros. Además, se explicará el fundamento teórico que justifica la selección de estos modelos y arquitecturas, así como las adaptaciones realizadas para adecuarlos a las particularidades de los datos en estudio.

4.3.1. Arquitectura para Datos Unidimensionales (1D, ver anexo A)

En este apartado, se detalla la arquitectura de la CNN diseñada específicamente para el análisis de datos en el dominio de la frecuencia. La arquitectura de la CNN incluye varios componentes esenciales, como capas convolucionales, capas de agrupamiento (pooling) y capas completamente conectadas. Cada una de estas capas contribuye al objetivo final de extraer características significativas de los datos y clasificarlos de la manera más precisa posible.

1. **Capa de Entrada:** La capa de entrada está configurada para recibir tensores con dimensiones de (8000,1). El primer valor, 8000, indica la cantidad de muestras contenidas en cada tensor. Por ejemplo, si se considera un fragmento de audio de 1 segundo muestreado a 8,000 Hz, dicho tensor contendrá 8,000 muestras correspondientes a ese segundo de audio. Cada una de estas muestras representa una medida puntual de la amplitud de la señal de audio en un instante específico. El segundo valor, 1, señala que el audio es monofónico en lugar de estereofónico; en el caso de audio estéreo, este valor sería 2. Esto indica que en un archivo de audio monofónico hay solo un canal, lo que significa que para cada instante temporal existe un único valor de amplitud.

2. **Bloques Conv1D (conv1d15, conv1d16, conv1d17 y conv1d14):** Cada uno de estas capas o bloques está estructurado con capas convolucionales unidimensionales (1D), las cuales se encuentran inmediatamente seguidas por capas de activación. Los filtros empleados en estas capas convolucionales tienen un tamaño de 128. Es pertinente mencionar que, para las capas conv1d15 y conv1d14, el número de parámetros es más bajo, siendo 512 y 256 respectivamente. Sin embargo, para las capas conv1d16 y conv1d17, el número de parámetros es de 49,280.

3. **Capas de activación:** Seguido de cada capa convolucional se aplica una función de activación lineal rectificada ReLu.

4. **Capa Residual o de Conexión de Residuales (add 4):** La capa de conexión residual, también conocida como .ªdd4", está diseñada para combinar directamente la salida de las capas conv1d17 y conv1d14. Esta combinación es una suma element-wise, lo que significa que cada elemento de la salida de conv1d17 se suma con el correspondiente elemento de la salida de conv1d14. Esta técnica se basa en la idea de que al combinar la información de estas dos capas, se retiene tanto la información original del input (conv1d14 que actúa como una capa de "shortcut") como las características más profundas detectadas por las capas intermedias (conv1d17). Esta conexión directa tiene el objetivo de mejorar el flujo de gradientes a lo largo de la red, previniendo problemas comunes como el desvanecimiento del gradiente. La razón por la que se eligen específicamente estas capas es para asegurar que tanto las características iniciales como las transformaciones más complejas sean consideradas en las capas posteriores.

5. **Capa pooling:** La red utiliza dos tipos de operaciones de pooling: Max-pooling y Average-pooling. La capa maxpooling1d4 aplica Max-pooling, que realza características específicas de las señales al retener el valor máximo de un conjunto determinado de entradas. Esto puede ayudar a identificar patrones o características distintivas en las señales analizadas.

6. **Capa de Pooling Promedio:** La capa average-pooling1d aplica Average-pooling, que calcula el promedio de un conjunto determinado de entradas. Esto es útil para suavizar las características y reducir la variabilidad dentro de las señales. Estas operaciones de pooling contribuyen a la reducción de dimensiones y la compresión de la información, permitiendo que la red se centre en características más generales y menos en detalles finos.

7. **Capa de aplanamiento:** La capa flatten transforma el tensor de salida de la capa averagepooling1d, que tiene dimensiones de (1333, 128), en un vector unidimensional con una longitud de 170,624 elementos. Al hacerlo, simplifica la estructura de los datos y prepara la información para ser procesada por las capas densas subsiguientes. Esta transformación es esencial para conectar capas convolucionales con capas totalmente conectadas, ya que estas últimas esperan entradas en forma de vectores. Al aplanar los datos, la red puede integrar y clasificar la información contenida en el tensor de manera más eficiente.

8. **Capa conectada completamente:** En estas capas, cada neurona está conectada a cada neurona de la capa anterior, y cada conexión tiene su propio peso. Estas conexiones permiten que la red aprenda relaciones complejas y patrones en los datos. La salida de una capa densa se calcula tomando el producto escalar del vector de entrada y la matriz de peso de la capa, agregando un término de sesgo y luego aplicando una función de activación. En la arquitectura utilizada, después de la capa de aplanamiento flatten, hay una serie de capas densas. La primera, dense, tiene 256 neuronas; la siguiente, dense1, tiene 128 neuronas; y finalmente, la capa output que tiene 5 neuronas, que corresponden a 5 clases diferentes para clasificación (en el primer experimeto que contepla solo cinco hablantes diferentes. Este parametro ira cambiando con cada uno de los experimentos propuestos). Estas capas densas son cruciales para el proceso de decisión y clasificación de la red.

4.3.2. Arquitectura para Datos Bidimensionales (2D, ver anexo A)

Para el análisis de espectrogramas en este estudio, se emplea una arquitectura de procesamiento de imágenes bidimensionales (2D) que se fundamenta en la estructura de la Red Neuronal Convolucional (CNN) conocida como AlexNet.

A diferencia de la arquitectura diseñada para el procesamiento de datos unidimensionales (1D), la CNN propuesta para datos bidimensionales (2D) requiere una fase de preprocesamiento de los datos. En esta fase se aplicó previamente un preprocesamiento utilizando la Transformada de Fourier de Tiempo Corto (STFT, por sus siglas en inglés) a cada uno de los archivos de audio en el conjunto de datos. Cabe destacar que la descripción que se proporcionará a continuación omite los detalles relacionados con el cálculo de los espectrogramas, centrándose exclusivamente en la arquitectura de la CNN.

1. **Capas Convolucionales:** La arquitectura inicia con dos capas convolucionales. La primera capa utiliza 32 filtros y 'kernels' de dimensiones (3,3). La segunda capa emplea 64 filtros, también con 'kernels' de dimensiones (3,3). Ambas capas tienen un 'stride' o paso de (1,1), lo que garantiza un desplazamiento uniforme a lo largo de las imágenes y captura de características a diferentes niveles.

2. **Capas de activación:** En este modelo, inmediatamente después de cada capa convolucional, se emplea la función de activación ReLU.

3. **Capas de pooling:** Después de cada capa convolucional, se añade una capa de Max Pooling. Estas capas emplean una ventana de (2,2) y reducen a la mitad las dimensiones espaciales del mapa de características, resaltando las características más prominentes y reduciendo la complejidad computacional.

4. **Capa de aplanamiento:** Antes de las capas densas o completamente conectadas, el modelo introduce una capa de aplanamiento (Flatten), que convierte los mapas de características 2D en vectores 1D.

5. **Capas completamente conectadas y de clasificación:** El modelo cuenta con dos capas densas. La primera tiene 64 neuronas y utiliza la función de activación ReLU. Entre estas dos capas densas, se encuentra una capa de "dropout"que ayuda a evitar el sobreajuste. Finalmente, la última capa densa, encargada de la clasificación, tiene tantas neuronas como clases a clasificar (en este caso, 6), y utiliza la función de cada 'softmax' para obtener las probabilidades de pertenencia a cada clase.

4.4. Arquitectura del Sistema Analítico para el Sistema de Interceptación de Comunicaciones

Una vez delineada la metodología para el procesamiento de señales de sonido, se procede a describir el sistema de análisis de datos que se propone implementar en el sistema de interceptación de comunicaciones de la FGN. Este proceso tiene como finalidad mejorar la eficiencia y eficacia en el uso de la información de alto valor probatorio que se produce en las salas de interceptación. El fin último de este sistema es identificar por medio del análisis de voz, si una nueva señal, es decir una nueva voz, ha sido o está siendo monitoreado en el sistema de interceptación Esperanza. Esta herramienta analítica facilitará los siguientes aspectos:

- **Asociación de casos:** Al identificar si una misma persona, utilizando distintas líneas telefónicas, está siendo monitoreada simultáneamente por dos o más salas, podrá reasignarse la carga de trabajo investigativa hacia una sola investigación. Esto permitirá una mayor concentración de recursos, lo que se traduce en una investigación más sólida y con mayor certeza probatoria.

- **Ubicación de la información:** Al introducir una nueva señal para análisis en el sistema, se puede determinar si esta ya ha sido monitoreada previamente. En caso afirmativo, el sistema es capaz de proporcionar los números de casos y las fiscalías que tienen procesos relacionados con

dicha señal.

- **Recolección de información:** Gracias a la capacidad del sistema para rastrear los procesos en los que está presente la señal analizada, es posible hacer solicitudes de inspección a aquellos procesos donde dicha señal ha sido identificada. Esto posibilita la recopilación de evidencia de alto valor probatorio.

- **Contextos más amplios:** Dado que el sistema se especializa en la ubicación y la asociación de señales de voz, posee una capacidad única de relacionamiento de interceptaciones telefónicas entre todos los interlocutores. Esta característica no solo facilita el monitoreo de las comunicaciones individuales, sino que también permite trazar una red más amplia de relaciones y conexiones, brindando una perspectiva holística que es indispensable para investigaciones complejas. De este modo, se pueden llevar a cabo investigaciones en contextos más amplios y se tiene acceso a una mayor cantidad de información, lo que permite una comprensión más profunda y matizada de las situaciones y facilita la toma de decisiones informadas basadas en evidencia sólida.

El sistema propuesto se estructura en tres procesos interrelacionados (ver figura 9). El primer proceso engloba la fase de registro, en la que se recolectan las señales de voz registradas en el sistema y se etiquetan con un número de registro único (registro de hash). Durante esta fase, se lleva a cabo la obtención de los espectrogramas o señales a partir de la información recolectada, incluyendo el preprocesamiento necesario para potenciar la precisión en el reconocimiento y la clasificación de las señales. Posteriormente, se procede al entrenamiento de la red neuronal convolucional, durante el cual se determinan los pesos que minimizan la función de costo y maximizan la presición del predicción.

El segundo proceso está orientado hacia el usuario final. Cuando este recibe una nueva señal y la introduce en el sistema, la información es procesada y se verifica si esta nueva señal corresponde a las que el modelo de aprendizaje ha sido entrenado para reconocer. En caso de coincidencia, el modelo emitirá su predicción más acertado y esta información será almacenada en la base de datos junto con los metadatos de la señal procesada.

Finalmente, en el tercer proceso se lleva a cabo la socialización de la información generada por el sistema. Aquí, se elaboran informes de análisis integrales que sintetizan los resultados obtenidos, facilitando así el proceso de toma de decisiones por parte de los fiscales.

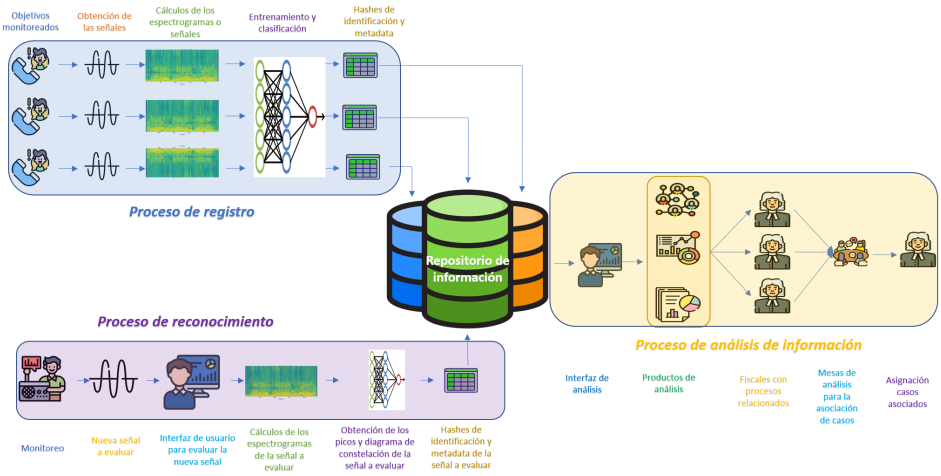


Figura 9: Proceso para el análisis, clasificación y relacionamiento
Elaboración propia.

Como se observa, el proceso demanda actualizaciones frecuentes. Esto implica que el modelo propuesto debe someterse a reentrenamientos periódicos para garantizar el procesamiento y predicción de nuevas señales. Aunque el escenario óptimo sería que el sistema se reentrenara de forma inmediata tras la identificación de cualquier nueva señal de interes investigativo, es viable determinar un umbral de señales no detectadas y ausentes en el repositorio de datos. La definición de estos umbrales y criterios debe ser establecida con base en análisis de coste-beneficio y consideraciones relativas a la seguridad de la información.

5. Experimentación y Resultados

La presente sección se dedica a la aplicación de los modelos teóricos descritos en las secciones anteriores. El principal objetivo de este ejercicio empírico es verificar la eficacia y precisión de los modelos al ser aplicados en contextos reales y proceder con la evaluación de los resultados derivados de dicha aplicación. En este marco, se presentará de forma detallada el diseño experimental empleado, se describirá el procedimiento de análisis adoptado, y se discutirán e interpretarán los resultados obtenidos. Es relevante destacar que este proceso no solo proporcionará una validación de las propuestas teóricas, sino que también permitirá identificar oportunidades de mejora, esenciales para futuras investigaciones y ajustes metodológicos. Durante la fase experimental, se evaluaron diversas métricas de desempeño para determinar la robustez del modelo propuesto, entre las que se incluyen indicadores como exactitud, precisión, sensibilidad (también conocida como *recall*) y la puntuación F1 (F1 score).

La metodología experimental adoptada consta de una serie de simulaciones que se llevan a cabo utilizando conjuntos de datos con atributos específicos. Estos conjuntos varían en términos de cantidad de hablantes y la duración de las grabaciones de voz.

El primer experimento se enfoca en la profundidad de los datos más que en su amplitud. Se basa en un conjunto de datos con un número reducido de hablantes, específicamente cinco, pero con una duración prolongada de registro vocal para cada uno, siendo 483 segundos. Esta configuración permite analizar en detalle las características vocales de cada hablante durante un periodo extenso.

El segundo experimento amplía el espectro de hablantes a seis, aunque se reduce la duración de las grabaciones a 400 segundos por hablante. Este enfoque equilibra la representatividad del conjunto de datos al abarcar más voces, pero con un análisis menos profundo en comparación con el primer experimento.

El tercer experimento sigue una tendencia similar, incorporando audios de trece hablantes distintos, pero cada grabación tiene una duración de 304 segundos. Esta configuración busca diversificar aún más las características vocales en el conjunto de datos.

Finalmente, el cuarto y último experimento se caracteriza por su enfoque en la diversidad. Incluye audios de 30 hablantes diferentes, con grabaciones de 103 segundos de duración para cada uno. Este diseño prioriza la representatividad y diversidad del conjunto de datos al capturar un amplio rango de características vocales distintas en un periodo de tiempo más corto.

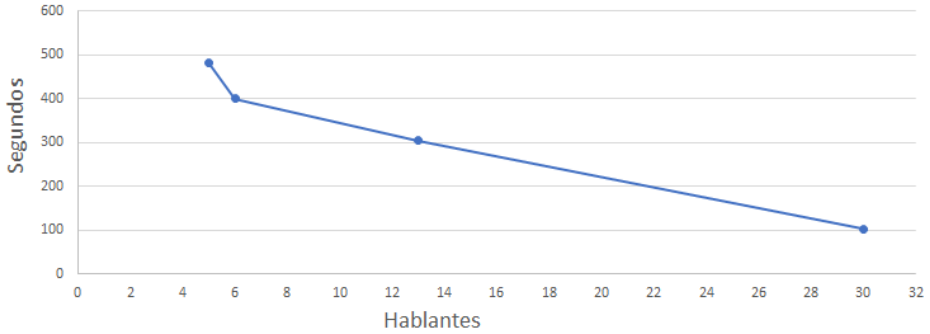


Figura 10: Relación duración de audio y número de hablantes
Fuente: Elaboración propia

5.1. Resultados

A continuación se muestran los resultados derivados del entrenamiento y predicción de los modelos. Para evaluar y contrastar el rendimiento entre los distintos modelos y experimentos propuestos, se han empleado las siguientes medidas de desempeño:

- **Pérdida**: La médida de pérdida se refiere a un valor que representa la diferencia entre las predicciones de un modelo y los datos reales. Es una medida cuantificable de qué tan lejos están los resultados de un modelo de los resultados esperados. Durante el proceso de entrenamiento, el objetivo suele ser minimizar esta pérdida, refinando el modelo para producir predicciones lo más cercanas posible a los valores reales. A diferencia de la precisión, no es un porcentaje sino una suma de errores para cada muestra. En el caso de las redes neuronales, la función de pérdida es el negativo del logaritmo de la función de verosimilitud. Entonces, naturalmente, el objetivo principal en un modelo de aprendizaje es reducir (minimizar) el valor de la función de pérdida con respecto a los parámetros del modelo cambiando los valores del vector de peso a través de diferentes métodos de optimización, como la retropropagación (backpropagation) en redes neuronales.

- **Exactitud:** La exactitud es una métrica utilizada para determinar la proporción de predicciones correctas realizadas por un modelo respecto del total de predicciones. Específicamente, la exactitud, cuantifica con qué frecuencia las etiquetas predichas del modelo coinciden con las etiquetas reales. Expresada como porcentaje, una exactitud del 100 % significa que las predicciones del modelo siempre son correctas, mientras que una exactitud del 0 % indica que todas las predicciones son incorrectas. $Exactitud = \frac{\text{Verdaderos positivos+Falsos positivos}}{\text{Verdaderos positivos}}$

- **Precisión:** La precisión es una métrica de rendimiento que evalúa la exactitud de las predicciones positivas realizadas por un modelo de clasificación. Específicamente, cuantifica la proporción de casos positivos previstos que se clasificaron correctamente. Este procedimiento se hace para cada una de las clases contempladas en el modelo, por lo que la medida de la precisión es en realidad un promedio de todas las clases contempladas. $Precision = \frac{\text{verdaderos positivos}}{\text{verdaderos positivos+falsos positivos}}$

- **Sensibilidad (Recall):**La sensibilidad, también conocida como recall, mide la proporción de positivos verdaderos que fueron correctamente identificados por el modelo. Específicamente, cuantifica cuántas de las instancias que realmente eran positivas en el conjunto de datos se identificaron correctamente en las predicciones del modelo. $Recall = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos positivos+Falsos negativos}}$

- **Puntuación F1:** La puntuación F1 es una métrica que combina tanto la precisión como la sensibilidad para proporcionar una valoración global del rendimiento de un modelo de clasificación. Esta puntuación brinda una manera de evaluar el equilibrio entre estas dos métricas críticas, reflejando en un único valor las compensaciones entre falsos positivos y falsos negativos. Dado que no siempre es suficiente evaluar un modelo solo en términos de precisión o solo en términos de sensibilidad, la puntuación F1 emerge como una medida valiosa, especialmente en escenarios donde ambos aspectos son importantes. $F1 \text{ Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$

En los cuatro experimentos, el Modelo 1 mostró consistentemente un mejor rendimiento que el Modelo 2, especialmente en las métricas de pérdida y exactitud, tal y como lo corroboran los valores obtenidos en estas mediciones. A pesar de que ambos modelos evidenciaron un incremento en la métrica de pérdida a lo largo de los experimentos, lo cual sugiere un incremento en la complejidad de las tareas de predicción, la discrepancia en su desempeño se hizo particularmente notoria en el cuarto experimento. Esta ampliación en la brecha podría deberse a desafíos surgidos por la inclusión de un mayor número de hablantes y la disminución de datos por hablante en las etapas posteriores de experimentación.

Cuadro 3: Metricas de Pérdida y exactitud a través de los experimentos y modelos evaluados

	Pérdida		Exactitud	
	Modelo 1	Modelo 2	Modelo 1	Modelo 2
Experimento 1	0.139	1.02	95.02 %	83.23 %
Experimento 2	0.198	1.09	93.55 %	83.90 %
Experimento 3	0.355	0.88	90.89 %	76.11 %
Experimento 4	0.857	2.46	84.07 %	36.14 %

Fuente: Cálculos propios

Como se puede apreciar en el panel (b) de la gráfica adjunta, la métrica de exactitud disminuye para

ambos modelos conforme se incrementa la complejidad del mismo. Sin embargo, es notable que el descenso en la exactitud del Modelo 2 es mucho más pronunciado en comparación con el Modelo 1. Esta notable diferencia resalta la robustez y versatilidad del Modelo 1 al enfrentarse a condiciones variadas y complejas.

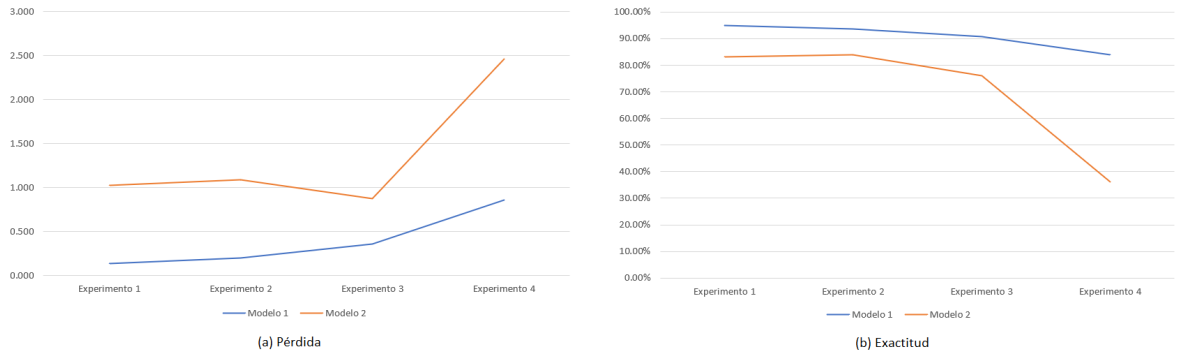


Figura 11: Métricas de desempeño de pérdida y exactitud para cada uno de los experimentos.
Fuente: Elaboración propia

A diferencia de lo observado con las métricas de pérdida y exactitud, las medidas de precisión, sensibilidad (recall) y puntuación F1 mostraron una tendencia diferente. A lo largo de los experimentos realizados, el Modelo 2 manifestó de manera consistente un rendimiento superior al del Modelo 1 en estas tres métricas: Precisión, Sensibilidad y Puntuación F1. El Modelo 1 exhibió un declive en su rendimiento en cada experimento, destacando una caída significativa en precisión, que descendió hasta un valor de 0,03. En contraste, el Modelo 2 sostuvo cifras destacadas, con niveles de precisión generalmente por encima de 0,77 y puntuaciones F1 que llegaron hasta 0,84. Si bien se identificó un valor atípico en la métrica de sensibilidad para el Modelo 1 en el primer experimento, los resultados subsiguientes reafirmaron la superioridad del Modelo 2 en identificar muestras positivas con una precisión más alta. En base a la evidencia presentada, el Modelo 2 se perfila como el más confiable y robusto para el objetivo propuesto.

Cuadro 4: Métricas de precisión, sensibilidad y puntuación F1 a través de los experimentos y modelos evaluados

	Precision		Recall		f1-score	
	Modelo 1	Modelo 2	Modelo 1	Modelo 2	Modelo 1	Modelo 2
Experimento 1	0.2	0.84	21	0.84	0.2	0.83
Experimento 2	0.18	0.84	0.18	0.84	0.18	0.84
Experimento 3	0.1	0.77	0.1	0.76	0.1	0.76
Experimento 4	0.03	0.39	0.02	0.38	0.03	0.34

Fuente: Cálculos propios

Como se observa el Modelo 1 ha mostrado una variabilidad significativa y una tendencia descendente en las tres métricas, el Modelo 2 ha demostrado ser más estable y con un rendimiento generalmente superior, lo que lo posiciona como el modelo más robusto.

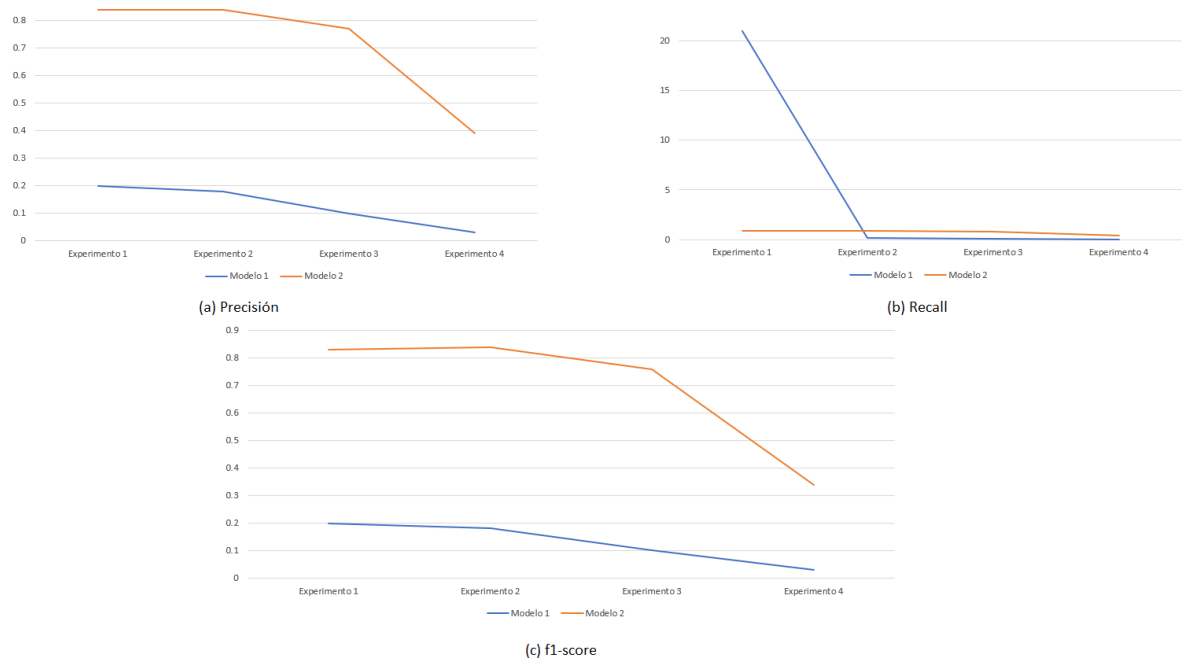


Figura 12: Métricas de precisión, sensibilidad y puntuación F1 a través de los experimentos y modelos evaluados.
Fuente: Elaboración propia

6. Conclusiones y futuras investigaciones[1]

La creciente necesidad de herramientas efectivas para el análisis de sonido en el campo judicial ha llevado al desarrollo presentado en esta tésis. Se logró diseñar con éxito una herramienta que permite la identificación de hablantes a partir de grabaciones de voz, asociándolos con registros previos en una base de datos. Adicionalmente, se diseño un flujo de trabajo que describe el proceso para implementar la herramienta diseñada en esta tesis y como esta puede impactar las investigaciones judiciales. Esta valiosa herramienta, con un enfoque principal hacia su aplicación en las salas de interceptación de la FGN, se perfila como un recurso altamente valioso para incrementar la eficiencia y precisión del sistema judicial. Su adopción institucional fortalecerá significativamente el acervo probatorio, facilitará la consolidación de casos judiciales más complejos y con un impacto más profundo, y optimizará los tiempos en los procesos de investigación. Un aspecto no menos importante es su potencial para reforzar la protección de los derechos fundamentales y la adecuada gestión de recursos investigativos.

[1]La redacción de las conclusiones de esta tesis fueron asistidas, a modo de experimento, con el modelo generativo Chatgpt 4 de Open IA.

No obstante, es importante señalar que, a pesar de que los modelos propuestos satisfacen adecuadamente los requisitos iniciales, existe un espacio evidente para su refinamiento. Los datos arrojados por el entrenamiento de los modelos destacan ciertas incongruencias entre las métricas de desempeño utilizadas. Mientras que indicadores tradicionales en el aprendizaje profundo, tales como la pérdida y exactitud, insinúan un mejor desepeño del Modelo 1, métricas cruciales como precisión, sensibilidad (recall) y puntuación F1 sugieren que el Modelo 2 presenta una solidez y coherencia superiores.

De cara al futuro, es necesario mantener un enfoque investigativo y proactivo hacia el perfeccionamiento de estos modelos. El avance de nuevos modelos o herramientas abren la oportunidad para explorar arquitecturas de redes neuronales más avanzadas y la posibilidad de integrar otros modelos y herramientas de inteligencia artificial que potencien aún más las capacidades actuales. El objetivo no es simplemente mejorar, sino alcanzar una precisión y confiabilidad que sean de plena confianza por los usuarios de este tipo de tecnologías, convirtiendo a este sistema en un referente en el ámbito judicial. Es crucial reconocer que los experimentos realizados ofrecen un entendimiento profundo sobre las demandas y requerimientos de un sistema de esta magnitud, anticipando una infraestructura que podrá albergar un volumen masivo de registros auditivos. Los resultados obtenidos hasta este punto permiten inferir que hay un amplio campo para futuras implementaciones a escala institucional y asegurando una preparación adecuada en términos de recursos, almacenamiento y capacidad de procesamiento.

7. REFERENCIAS

[1] C. Macleoad, *abracadabra: How does Shazam work?* https://www.cameronmacleod.com/blog/how-does-shazam-work, 2022.

[2] F. G. de la Nación, "Directiva 0004," *Fiscalía General de la Nación*, 2021.

[3] Fiscalia General de la Nación, "Directiva 004 de 2021," 2021, 02 de noviembre de 2021.

[4] R. S. U. Investigativa, "Esperanza: el misterioso sistema de interceptaciones del caso uribe- cepeda," *Revista Semana*, 2018.

[5] A. L.-C. Wang, "An industrial-strength audio search algorithm," *Shazam Entertainment, Ltd.*, 2013.

[6] H. Salehghaffari, "Speaker verification using convolutional neuronal networks," *Control Research Laboratory*, vol. 14, no. 3, pp. 342–351, 2018.

[7] A. K. M. S. M. B. M. Wang, T. Sirlapu and R. Nicolas, "Speaker recognition using convolutional neural network with minimal training data for smart home solutions," *International Conference on Human System Interaction (HSI)*, vol. 14, pp. 139–145, 2018.

[8] Y. Jia, X. Chen, J. Yu *et al.*, "Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network," *Complex Intelligent Systems*, vol. 7, pp. 1749–1757, 2021. [Online]. Available: https://doi.org/10.1007/s40747-020-00172-1

[9] N. K. J. Hourri, Soufiane. Nikolo, "Convolutional neural network vectors forspeaker recognition," *Vol.:(0123456789)1 3International Journal of Speech Technology*, vol. 24, pp. 389–400, 2021.

[10] S. R. Arshad, S. M. Haider, and A. B. Mughal, "Speaker identification using speech recognition," *ArXiv*, vol. abs/2205.14649, 2022.

[11] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," 2017.

[12] N. Dave, "Feature extraction methods lpc, plp and mfcc in speech recognition," *International Journal For Advance Research in Engineering And Technology(ISSN 2320-6802)*, vol. Volume 1, 07 2013.

[13] M.-W. Mak and J.-T. Chien, *Machine learning for speaker recognition.* Cambridge University Press, 2020.

A. Apéndice

Cuadro 5: Arquitecutra CNN para datos unidimensionales

Layer (type)	**Output Shape**	**Param #**	**Connected to**
input (InputLayer)	(None, 8000, 1)	0	-
conv1d_15 (Conv1D)	(None, 8000, 128)	512	input[0][0]
activation_10 (Activation)	(None, 8000, 128)	0	conv1d_15[0][0]
conv1d_16 (Conv1D)	(None, 8000, 128)	49280	activation_10[0][0]
activation_11 (Activation)	(None, 8000, 128)	0	conv1d_16[0][0]
conv1d_17 (Conv1D)	(None, 8000, 128)	49280	activation_11[0][0]
conv1d_14 (Conv1D)	(None, 8000, 128)	256	input[0][0]
add_4 (Add)	(None, 8000, 128)	0	conv1d_17[0][0], conv1d_14[0][0]
activation_12 (Activation)	(None, 8000, 128)	0	add_4[0][0]
max_pooling1d_4 (MaxPooling1D)	(None, 4000, 128)	0	activation_12[0][0]
average_pooling1d (AveragePooling1D)	(None, 1333, 128)	0	max_pooling1d_4[0][0]
flatten (Flatten)	(None, 170624)	0	average_pooling1d[0][0]
dense (Dense)	(None, 256)	4368000	flatten[0][0]
dense_1 (Dense)	(None, 128)	32896	dense[0][0]
output (Dense)	(None, 5)	645	dense_1[0][0]
Total params: 43812869 (167.13 MB)			
Trainable params: 43812869 (167.13 MB)			
Non-trainable params: 0 (0.00 Byte)			

Cuadro 6: Arquitectura de la red convolucional para datos 2d

Layer (type)	**Output Shape**	**Param #**
conv2d (Conv2D)	(None, 126, 43, 32)	320
max_pooling2d (MaxPooling2D)	(None, 63, 21, 32)	0
conv2d_1 (Conv2D)	(None, 61, 19, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 30, 9, 64)	0
flatten (Flatten)	(None, 17280)	0
dense (Dense)	(None, 64)	1105984
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 6)	390
Total params: 1125190 (4.29 MB)		
Trainable params: 1125190 (4.29 MB)		
Non-trainable params: 0 (0.00 Byte)		