Universidad Francisco de Vitoria

Grado en Ingeniería Matemática

Clasificador de expresiones auditivas en entornos sociales y de cuidado

Equipo:

Miguel Egido Morales, Jonás Manuel García Vallejo, Alejandra Llord Muñoz de la Espada, Alfredo Robledano Abasolo

Abril 2024

Índice

1. Resumen **1**

2. Introducción **1**
2.1. Objetivos 2
2.1.1. Objetivo general 2
2.1.2. Objetivos específicos 3
2.2. Alcance e impacto 3

3. Estado de la cuestión **4**
3.1. Investigaciones existentes 6
3.2. Tabla resumen 10

4. Resolución **12**
4.1. Desarrollo y proceso completo 12
4.1.1. Database 13
4.1.2. Preprocesamiento 13
4.1.3. Entrenamiento 15
4.1.4. Comparativa de modelos 16
4.1.5. Implementación 17

5. Conclusión **17**

1. Resumen

Este proyecto desarrolla un modelo de clasificación acústica no verbal que implementa diferentes técnicas de algoritmia, centrándose en el aprendizaje automático y las redes neuronales. La base de datos consiste en un conjunto de audios no verbales de distintos tipos (toses, suspiros, estornudos...) y la metodología implementada es KDD (Knowledge Discovery in Databases), que es muy eficaz con grandes volúmenes de datos. Finalmente, se ha logrado obtener un 86 % de precisión en las clasificaciones y una aplicación que implementa el clasificador en directo; pudiendo así comprobar la eficacia y posible implementación del modelo.

2. Introducción

En la era digital actual, la capacidad de las máquinas para interpretar y responder a nuestro entorno ha transformado diversas áreas de investigación y desarrollo tecnológico. Aquellas campos de investigación que aprovechan las posibilidades que brindan las nuevas tecnologías de inteligencia artificial y aprendizaje automático, están experimentando un crecimiento exponencial.

La clasificación y reconocimiento de sonidos, especialmente en contextos sociales y de salud, está siendo un campo de gran interés debido a sus amplias aplicaciones prácticas.

Desde la monitorización de condiciones en pacientes hasta la interacción mejorada entre humanos y dispositivos inteligentes. En la actualidad, casi la totalidad de dispositivos tecnológicos que salen al mercado incorporan asistentes de voz inteligentes capaces de entender el habla humana. Tomando en cuenta estos hechos, se puede concluir que el análisis automático de sonidos ofrece prometedoras avenidas para mejorar la calidad de vida y la eficiencia operativa en múltiples dominios.

Dentro de este área de interés, los sonidos humanos que clasificamos como "no-verbales" tales como la tos, las risas, los llantos y los suspiros son indicadores vitales de diversas condiciones; tanto anímicas como físicas y de salud . En entornos como hospitales y residencias de ancianos, la identificación precisa de estos sonidos puede ser esencial para la provisión de atención y bienestar que asegure unos cuidados de calidad a grupos de población más vulnerables. Sin embargo, la detección y clasificación eficaz de dichos sonidos en entornos ruidosos y no estructurados sigue presentando un desafío técnico significativo debido a diversos factores como las condiciones ambientales desfavorables y las limitaciones de software actuales.

Con lo cual, este proyecto propone desarrollar y utilizar algoritmos avanzados de aprendizaje automático para la clasificación precisa de sonidos específicos en diferentes entornos sociales para su posterior implementación. A través del uso de técnicas como el procesamiento de señales y redes neuronales profundas, se busca crear un sistema capaz de identificar de manera fiable sonidos humanos , haciendo una diferenciación clara de otros ruidos de fondo y asegurando una respuesta contextual apropiada.

Tras muchas consideraciones , se ha determinado que este proyecto podría ser de gran importancia, pues tiene un amplio potencial para impactar positivamente en la sociedad; especialmente en el ámbito de la salud. Por ejemplo, en hospitales, la detección automática de sonidos como la tos o el llanto puede ayudar al personal médico a monitorizar más efectivamente el estado de los pacientes, permitiendo intervenciones más rápidas y adecuadas. Asimismo, en otros ámbitos como el del hogar inteligente, este sistema puede mejorar la interacción entre los usuarios y sus entornos; facilitando un ambiente más reactivo y sensible a sus necesidades.Un ejemplo de ello sería el ajuste automático de la temperatura y humedad del hogar al detectar un elevado nivel de toses o estornudos. En conclusión, la identificación y clasificación de sonidos no verbales no solo es un desafío técnico, sino que también surge como una necesidad emergente que está trayendo avances significativos para la comodidad y el bienestar de la población.

2.1. Objetivos

2.1.1. Objetivo general

Identificar y extraer diferentes tipos de sonidos a partir de fragmentos de audio para su posterior análisis e implementación en diferentes modelos sociales. Para ello aplicaremos la Transformada de Fourier para limpiar el sonido de nuestros fragmentos de audio. Posteriormente implementaremos un modelo de clasificación, el cual entrenaremos y haremos predicciones con nuevos datos.

2.1.2. Objetivos específicos

1. **Explorar y adquirir datos de audio:** Recopilar una amplia variedad de muestras de audio que representen diferentes tipos de sonidos relevantes para el análisis y modelado social.
2. **Preprocesamiento de datos de audio:** Utilizar herramientas de procesamiento de señales para preprocesar los datos de audio, realizando la normalización del volumen y la eliminación de ruido de fondo no deseados. Este último lo realizaremos mediante la aplicación de la transformada de fourier para representar la señal en el dominio de la frecuencia, analizar este espectro de frecuencia para identificar las componentes de sonido deseadas y el ruido no deseado y por último realizar una eliminación selectiva de frecuencias no relevantes para mejorar la calidad de la señal y reducir el impacto del ruido.
3. **Implementar la Transformada de Fourier:** Utilizaremos la biblioteca *numpy* de Python, que permitirá transformar los datos de audio del dominio temporal al dominio de la frecuencia.
4. **Análisis de frecuencia de sonido:** Analizar los espectros de frecuencia resultantes para identificar las características espectrales distintivas de diferentes tipos de sonidos, como tonos, frecuencias dominantes y distribuciones de energía en diferentes bandas de frecuencia.
5. **Extracción de características relevantes:**Identificar y extraer características relevantes de los espectros de frecuencia que puedan ser útiles para la clasificación y análisis de diferentes tipos de sonidos, como la amplitud máxima, la frecuencia dominante o la dispersión espectral.
6. **Desarrollo de modelos de clasificación:**Utilizaremos técnicas de aprendizaje automático, como redes neuronales, para desarrollar un modelo que pueda clasificar automáticamente los diferentes tipos de sonidos basados en las características extraídas.
7. **Validación y evaluación del modelo:**Evaluaremos el rendimiento de los modelos de clasificación con nuevos datos para evaluar el rendimiento de nuestro modelo. Utilizando métricas como la precisión.
8. **Implementación y despliegue del modelo:**Integraremos el modelo de clasificación final en una aplicación para su implementación práctica (de manera únicamente teórica), lo que llegará a poder realizar una detección y clasificación a tiempo real en entornos sociales específicos.

2.2. Alcance e impacto

Nuestro equipo compuesto por cuatro ingenieros matemáticos se enfrenta al objetivo de desarrollar un modelo de aprendizaje automático, para clasificar audios con diferentes sonidos como tos, risa, estornudos... etc; dado que no disponemos de un presupuesto asignado, utilizaremos recursos gratuitos y códigos abiertos para llevar a cabo el

proyecto.

Los objetivos propuestos por el equipo del proyecto tratan del desarrollo de un modelo de aprendizaje automático, esto conllevará la recopilación de datos, la selección de características relevantes de los audios, el entrenamiento del modelo y la comprobación de su rendimiento, por último se hará un análisis de los resultados obtenidos para sacar conclusiones sobre las aplicaciones del modelo que puedan llegar a generar un impacto significativo, además identificaremos posibles mejoras para futuras iteraciones del proyecto.

Las fechas de los objetivos estarán bien definidas en el diagrama de Gantt, anteriormente hecho, que será utilizado para la planificación del progreso del proyecto. Este diagrama será una herramienta que verificará con eficiencia con el cumplimiento de los plazos establecidos y consigamos así un avance constante en todas las etapas del proyecto.

Nuestro proyecto tiene el potencial de generar impactos significativos para la sociedad, la tecnología, beneficios sociales y medioambientales. Al reconocer y detectar eventos de sonidos relevantes, podría aplicarse en mejoras en sistemas de seguridad y análisis de datos para sacar la información más relevante, un ejemplo sería en las llamadas de atención al cliente para comprobar si el cliente está satisfecho o no con la empresa, también podría ser empleado por médicos para realizar un seguimiento de sus pacientes extrayendo los sonidos relevantes y facilitando diagnósticos más precisos

En términos medioambientales nuestro proyecto podría tener un impacto significativo, por ejemplo, en la mejora de detección de sonidos, esto podría ayudar a la seguridad pública, al poder dar respuestas más rápidas y eficientes en situaciones de emergencia.

3. Estado de la cuestión

El análisis y procesamiento de señales de audio han experimentado avances significativos gracias a la aplicación de técnicas matemáticas y computacionales. En este contexto, la Transformada de Fourier y las Redes Neuronales emergen como herramientas esenciales, impulsando innovaciones en el reconocimiento de patrones de audio, el procesamiento de señales y la mejora de sistemas de diagnóstico.

Transformada de Fourier

La Transformada de Fourier juega un papel crucial en el desglose de señales de audio en sus componentes frecuenciales, facilitando el análisis y la clasificación detallada de las señales. Aguirre Martín (2017), en su trabajo sobre el desarrollo y análisis de clasificadores de señales de audio, destaca la importancia de esta herramienta para el procesamiento efectivo de señales de audio, especialmente en la identificación de características únicas dentro de señales complejas.

Redes Neuronales

Por otro lado, las Redes Neuronales, particularmente las Convolucionales (CNNs), han demostrado ser excepcionalmente eficaces en el reconocimiento y clasificación de señales de audio. Su capacidad para aprender de grandes cantidades de datos y extraer características significativas sin la necesidad de preprocesamiento detallado ha revolucionado el campo. Chachada y Kuo (2014) proporcionan una revisión exhaustiva sobre el reconocimiento de sonidos ambientales, resaltando cómo las técnicas basadas en aprendizaje automático han avanzado en la clasificación y el análisis de sonidos no musicales y no verbales.

Integración de ambas técnicas

La integración de la Transformada de Fourier con Redes Neuronales ha permitido el desarrollo de sistemas de reconocimiento de audio altamente efectivos y eficientes. Esto es debido a la capacidad de aplicar la Transformada de Fourier en entornos ruidosos y complejos para disminuir las interferencias y ampliar la frecuencia de interés. De esta forma, se pueden optimizar todos los datos independientemente del entorno y así aprovechar al máximo la capacidad de las redes neuronales para desarrollar diferentes modelos. Este enfoque combinado se ha aplicado en una variedad de aplicaciones, desde el diagnóstico médico hasta la seguridad y la interacción hombre-máquina.

KNN (K-Nearest Neighbors)

El algoritmo KNN (K-Nearest Neighbors) es una técnica de aprendizaje automático utilizada para clasificar objetos basándose en la similitud de sus características con las de los objetos cercanos en un espacio de características definido. Al asignar una etiqueta a un nuevo punto de datos, KNN examina las etiquetas de los puntos de datos vecinos más cercanos y asigna la etiqueta más común o promedio entre ellos.

Tensores

Los tensores son estructuras de datos multidimensionales que generalizan los conceptos de escalares, vectores y matrices a dimensiones superiores. Para el caso de procesamiento de señales, los tensores se utilizan para representar datos complejos, como imágenes, audio o series temporales. Su versatilidad permite realizar operaciones matemáticas eficientes y manipulaciones de datos en grandes conjuntos de datos.

Mel-frequency cepstral coefficients (MFCCs)

Los coeficientes cepstrales de frecuencia de Mel (MFCCs) son una técnica ampliamente utilizada en el procesamiento de señales de audio para extraer características importantes que describen la estructura espectral de una señal sonora. Los MFCCs se calculan mediante un proceso que simula la respuesta del oído humano a diferentes frecuencias, lo que los hace especialmente útiles para el reconocimiento y clasificación de sonidos en entornos ruidosos o con variaciones de frecuencia.

Aplicaciones destacadas

Deiagnóstico médico

La capacidad para clasificar y analizar señales de voz y sonidos cardíacos con precisión tiene implicaciones significativas en el campo médico, permitiendo el desarrollo de herramientas de diagnóstico no invasivas y altamente efectivas.

Seguridad y monitoreo ambiental

El reconocimiento preciso de sonidos ambientales puede ser utilizado para mejorar sistemas de seguridad y monitoreo, ofreciendo métodos rápidos y confiables para la detección de eventos inusuales o peligrosos.

Desafios y futuro

A pesar de los avances logrados, persisten desafíos significativos, particularmente en la mejora de la precisión del reconocimiento en entornos ruidosos y la gestión eficiente de grandes volúmenes de datos. La investigación futura se beneficiará de un enfoque multidisciplinario, combinando conocimientos de acústica, matemáticas, y ciencias de la computación, para desarrollar soluciones innovadoras que aborden estos desafíos.

3.1. Investigaciones existentes

Como aproximación inicial a la realización del proyecto se ha decidido buscar e investigar 8 artículos relacionados con el contexto del proyecto, tanto ejemplos prácticos como investigaciones y diferentes publicaciones teóricas/académicas. La finalidad de reunir estos 8 artículos es la de imitar y seguir buenas prácticas que en otros proyectos hayan implementado y que nos puedan ser de utilidad para entender el tipo problema que se va a resolver.

Artículo 1

En el artículo "Desarrollo y análisis de clasificadores de señales de audio"[1] podemos encontrar un análisis sobre la clasificación y reconocimiento de señales de audio, donde podemos destacar su relevancia en aplicaciones modernas que combinan procesamiento de audio y aprendizaje automático. En este artículo se mencionan temas bien establecidos como el reconocimiento de voz y la música, así como un campo emergente, la clasificación automática de ruidos ambientales. Para el aprendizaje automático emplea un modelo supervisado, donde los datos de audio están etiquetados en clases. Se describe un proceso de clasificación, que abarca desde la extracción de características del audio hasta la aplicación de algoritmos de aprendizaje automático para esta tarea. También, se proporciona un sistema de clasificación y etiquetado de audio en tiempo real como ejemplo concreto de aplicación práctica, incluyendo una estructura de clasificación jerárquica.

Esta información puede ser muy útil para orientar nuestro proyecto en el desarrollo de un modelo que diferencie audios de tos, estornudos y risas. También es importante para nuestro proyecto la extracción de características del audio y el uso de algoritmos de aprendizaje automático para la clasificación. Además, se menciona un enfoque supervisado, que es importante cuando tienes datos etiquetados en clases. Por último, el artículo menciona el desarrollo de un software para la clasificación de audio en tiempo real, lo cual puede ser útil para implementar el modelo en un entorno práctico.

Artículo 2

El artículo "Environmental sound recognition: a survey"[3] ofrece una revisión detallada y comprensible de los avances más recientes en el campo del reconocimiento de sonidos ambientales, organizada en cuatro secciones principales. Estas incluyen una explicación de los enfoques básicos de procesamiento de sonidos ambientales, técnicas para manejar sonidos ambientales estacionarios y no estacionarios, así como una comparación del rendimiento entre diferentes métodos. Por último, el artículo discute las conclusiones obtenidas y señala las tendencias futuras en la investigación y desarrollo de este campo específico.

Este artículo es útil porque proporciona revisión detallada de los avances de campo del reconocimiento de sonidos ambientales. Al revisar las técnicas presentadas en el artículo, podrás identificar técnicas relevantes que podrían ser aplicables a tu proyecto de diferenciar entre audios de tos y risa. Por ejemplo, podrías encontrar métodos para manejar aspectos no estacionarios de los sonidos, lo cual puede ser relevante para identificar características distintivas de la tos y la risa en los audios.

Artículo 3

La publicación "Clasificación automática de sonidos utilizando lenguaje máquina"[5] realiza la implementación de dos modelos para clasificar sonidos ambientales mediante aprendizaje automático. Ambos resaltan la importancia del procesamiento del espectrograma de audio y el preprocesamiento de datos. La finalidad de este artículo es crear un sistema para clasificar sonidos ambientales con precisión aceptable, explorando métodos de extracción de características y utilizando inteligencia artificial.

Este trabajo es útil ya que aborda un tema muy conectado al nuestro, nuestro objetivo es la clasificación de sonidos producidos por las personas. Aunque este artículo clasifique sonidos ambientales nos puede ayudar a relacionar el enfoque que podemos darle a nuestro trabajo.Además utiliza la transformada de fourier, luego también es útil para empezar a conocerla para poder trabajar con ella.

Artículo 4

Este proyecto aborda el reconocimiento de instrumentos musicales mediante múltiples algoritmos de aprendizaje automático, incluyendo KNN, redes neuronales, PCA, LDA y Random Forest. En la publicación "Reconocimiento automático de instrumentos mediante aprendizaje máquina"[7] se evalúan distintos enfoques y se concluye que, en

algunos casos, los algoritmos más simples, como KNN, ofrecen resultados óptimos. Se destaca la importancia de la reducción de dimensiones y se exploran las dificultades asociadas con instrumentos específicos.

La utilidad de la información del estudio de instrumentos se refleja en la experiencia adquirida en el procesamiento de datos de audio y en la selección de algoritmos eficaces para la clasificación, realizando un contraste de estos para elegir el mejor candidato. Además, conceptos como la reducción de dimensiones y la adaptabilidad a contextos cambiantes pueden aplicarse al reconocimiento de sonidos en nuestro proyecto.

Artículo 5

En el estudio piloto "Classification of Nonverbal Human Produced Audio Events: A Pilot Study"[2], se plantea la necesidad de avanzar en el estado del arte sobre las capacidades de procesamiento en micro-controladores de sonidos no verbales como la tos y carraspeo. La correcta clasificación de este tipo de sonidos no verbales captados a través de un pequeño dispositivo intra-aural podría mejorar la atención médica. Las técnicas propuestas que se evaluan en este estudio piloto son Gaussian Mixture Model (GMM), Support Vector Machine y Multi-Layer Perceptron.

El estudio trata técnicas que se adecuan enormemente al tipo de problema que se quiere resolver. Entender su funcionamiento y complejidad permite aumentar el abánico de posibilidades en la tarea de desarrollar un modelo clasificador de sonidos no verbales.

Artículo 6

En la publicación "A Bag-of-Audio-Words Approach for Snore Sounds' Excitation Localisation"[8] se estudia como mejorar la calidad del sueño y tratar patologías graves en personas que padecen de ronquidos y de apnea del sueño, a partir del análisis de sonidos. Se trata de un campo de estudio eminentemente práctico que destaca por la importancia y los riesgos en la salud que generarn este tipo de enfermedades. Se utilizan técnicas como Bag-of-audio-words para mejorar la eficiencia en clasificadores de sonidos.

El enfoque que se da al análisis de sonidos es muy práctico y aunque tal vez muy específico, permite hacerse a la idea de como puede ser una aplicación práctica de un clasificador de sonidos, además de mejorar la precisión con la técnica Bag-of-audio-words.

Artículo 7

El artículo "Eliminacion de ruido en sonidos cardíacos mediante tecnicas de aprendizaje profundo"[6] aborda la problemática del ruido en grabaciones de sonidos cardiacos y propone una solución mediante el uso de técnicas avanzadas de aprendizaje profundo. Se enfoca en mejorar la calidad de estas grabaciones para facilitar diagnósticos más precisos y eficaces en el ámbito médico. La investigación destaca por su enfoque innovador

en el procesamiento de señales y su potencial para contribuir significativamente a la cardiología y la medicina diagnóstica.

Artículo 8

El objetivo principal del proyecto "Clasificación de voces a través de séries de Fourier y redes neuronales"[4] es desarrollar un método matemático que permita clasificar las voces según si padecen algún tipo de afección o no, utilizando para ello series de Fourier y redes neuronales convolucionales (CNN). El trabajo se divide en dos grandes bloques: la clasificación mediante series de Fourier y la clasificación utilizando redes neuronales. En el primer bloque, se escogen voces sanas y voces con nódulos para comparar su frecuencia fundamental media y la amplitud de sus oscilogramas. Se plantean hipótesis sobre las diferencias entre voces sanas y con patología, proponiendo ajustar distribuciones gaussianas a las características de las señales de voz para confirmar o deshechar estas hipótesis.
En el segundo bloque, se propone el uso de redes neuronales, específicamente la red neuronal convolucional (CNN), para la clasificación de las voces. Se identifica un problema de desbalance en el conjunto de datos, ya que hay muchas más muestras de voces sanas que de voces con patología. A pesar de que la exactitud del modelo es alta, la precisión y el recall son muy bajos para las muestras patológicas, lo que indica que el modelo tiende a clasificar todas las señales de voz como sanas. Para intentar solucionar este problema, se propone un entrenamiento de bajo y sobremuestreo de las entradas de la red.

El manejo de redes neuronales en sintonía con la transformada de Fourier es el foco principal del proyecto. Estas herramientas permitirán una correcta limpieza y extracción de características del dataset de sonidos no verbales.

3.2. Tabla resumen

Artículo	Título	Análisis	Objetivo del estudio	Relación y aportación
1	Desarrollo y análisis de clasificadores de señales de audio	Crear un sistema de clasificación de sonidos ambientales.Una vez implantado el modelo estudiar los parámetros de decisión.	Explorar métodos de extracción de características de audio.Conseguir una precisión aceptable. Elegir el mejor modelo de inteligencia artificial.	Clasificador de sonidos y Transformada de Fourier
2	Environmental sound recognition: a survey	Desarrollar métodos para el reconocimiento de instrumentos musicales utilizando algoritmos de aprendizaje automático	Identificar qué algoritmo es más óptimo para la clasificación. Se exploran aspectos como la reducción de dimensiones y las dificultades asociadas a cada instrumento	Contraste de algoritmos de clasificación y analogía de los instrumentos musicales con sonidos producidos por las personas.
3	Clasificación automática de sonidos utilizando lenguaje máquina	Crear un sistema de clasificación de sonidos ambientales.Una vez implantado el modelo estudiar los parámetros de decisión.	Explorar métodos de extracción de características de audio.Conseguir una precisión aceptable. Elegir el mejor modelo de inteligencia artificial.	Clasificador de sonidos y Transformada de Fourier

4	Reconocimiento automático de instrumentos mediante aprendizaje máquina	Desarrollar métodos para el reconocimiento de instrumentos musicales utilizando algoritmos de aprendizaje automático	Identificar qué algoritmo es más óptimo para la clasificación. Se exploran aspectos como la reducción de dimensiones y las dificultades asociadas a cada instrumento	Contraste de algoritmos de clasificación y analogía de los instrumentos musicales con sonidos producidos por las personas.
5	Classification of Nonverbal Human Produced Audio Events: A Pilot Study	Evaluación de la técnicas GMM, SVM y Perceptrón Multicapa en el análisis de sonidos no verbales.	En 10 categorías de sonidos no verbales, medir la eficiencia de 3 técnicas comunmente utilizadas en el procesamiento de sonidos.	Guía útil en la elección del algoritmo utilizado en el desarrollo de un clasificador de sonidos no verbales.
6	A Bag-of-Audio-Words Approach for Snore Sounds' Excitation Localisation	Análisis de sonidos relacionados con el sueño (ronquidos, apnea del sueño, ...).	Realizar un enfoque en la clasificación basado en la técnica Bag-of-Audio-Words.	Entender un caso práctico de uso y el enfoque mediante la técnica Bag-of-Audio-Words.
7	Eliminacion de ruido en sonidos cardíacos mediante tecnicas de aprendizaje profundo	Estudia la efectividad de las técnicas de aprendizaje profundo para eliminar el ruido de las grabaciones de sonidos cardíacos.	El estudio busca mejorar la calidad de los registros de sonidos cardíacos, lo que permitiría un diagnóstico más preciso y confiable de las condiciones cardíacas.	Clasificador de sonidos, técnicas avanzadas de aprendizaje profundo y Transformada de Fourier

8	Clasificación de voces a través de séries de Fourier y redes neuronales	Se centra en el desarrollo y evaluación de un método matemático para clasificar voces según la presencia de alguna afección.	El estudio busca identificar diferencias significativas en las características de las señales de voz entre voces sanas y voces con patología, con el fin de proporcionar una herramienta precisa y eficaz para el diagnóstico temprano de trastornos vocales.	Series de Fourier y Redes Neuronales Convolucionales (CNN).

4. Resolución

4.1. Desarrollo y proceso completo

Para el desarrollo del proyecto, se ha optado por utilizar la metodología KDD (Knowledge Discovery in Databases) debido a su eficacia en proyectos de análisis y clasificación de datos. KDD proporciona una serie de pasos estructurados que abarcan desde la identificación de los datos hasta la implementación de modelos predictivos, esto es clave para manejar de forma eficaz la complejidad de nuestros datos de audio y garantizar una gestión coherente y organizada de todo el proyecto. Además, KDD ofrece un enfoque iterativo que nos permite ajustar y mejorar continuamente nuestros modelos a medida que avanzamos en el análisis de los datos.

En cada una de las cinco etapas de la metodología KDD, se llevan a cabo procesos específicos para avanzar en el análisis y la clasificación de los datos de audio. La primera etapa, denominada "database", se centra en la identificación y selección de los datos relevantes para nuestro análisis. Luego, en la etapa de "preprocesamiento", se realizan diversas operaciones, para preparar los datos de manera óptima para su análisis posterior.

La siguiente etapa es el "entrenamiento", donde se desarrollan y ajustan los modelos de clasificación de sonidos utilizando técnicas de aprendizaje automático. Posteriormente, en la etapa de "comparativa de modelos", se evalúan y comparan los diferentes modelos desarrollados para identificar el más adecuado para nuestro propósito. Finalmente, en la etapa de "implementación", se llevan los modelos seleccionados a entornos prácticos,

lo que nos permite utilizar los resultados obtenidos en situaciones del mundo real.

4.1.1. Database

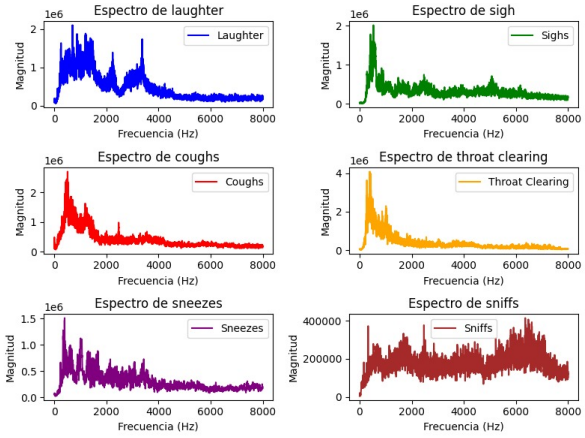
Comenzando por la parte de database, se realizó un análisis inicial de los datos disponibles. Se procesaron un total de 2,66 GB de datos, que consistían en 21,024 fragmentos de audio con diferentes tipos de sonidos no verbales de longitud variable. Sin embargo, durante este proceso se realizó una limpieza de los datos para eliminar fragmentos de audio con longitudes similares a 0, lo que resultó en un conjunto final de 20,502 audios para su análisis

4.1.2. Preprocesamiento

En cuanto al preprocesamiento de los datos, se llevaron a cabo varias etapas importantes. En primer lugar, se realizó la normalización de los archivos de audio en formato ".wav", transformándolos en tensores de valores decimales. Además, se asignaron etiquetas en formato one-hot a los nombres de los archivos para facilitar su procesamiento y clasificación.

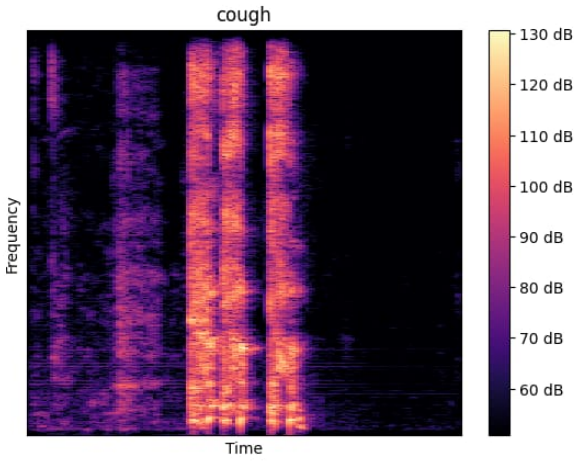
Posteriormente, se llevo a cabo la estandarización de las señales de audio para igualar las longitudes de los tensores de señal. Esto se logró mediante dos técnicas principales: el padding, que consiste en agregar valores nulos a los extremos de los tensores para igualar su longitud, y el random cropping, que recorta aleatoriamente fragmentos de los tensores para ajustar su longitud.

Además, se exploraron diversas técnicas para el procesamiento y análisis de datos de audio. Inicialmente, se optó por aplicar la transformada de Fourier a los fragmentos de audio. Esta técnica permitió descomponer las señales de audio en sus componentes de frecuencia, lo que resultó útil para identificar las características espectrales distintivas de diferentes tipos de sonidos. Sin embargo, conforme se avanzaba en el proyecto, se observaron limitaciones en términos de optimización y eficacia en la clasificación de sonidos al aplicar exclusivamente la transformada de Fourier.



Por lo tanto, se decidió explorar otras técnicas y se investigó Mel-frequency cepstral coefficients (MFCCs). Los Mel-frequency cepstral coefficients (MFCC) son una técnica

comúnmente usada en el procesamiento de señales de audio para extraer características importantes que describen la estructura espectral de una señal sonora. Esta técnica se basa en la idea de que el sistema auditivo humano no percibe todas las frecuencias de manera uniforme, sino que es más sensible a ciertos rangos de frecuencia, lo que se conoce como la escala de Mel.



No obstante, se continuó en la búsqueda por la mejor estrategia de procesamiento de datos. Después de implementar los MFCCs, se buscó la opinión de un consultor ingeniero externo. Este paso resultó crucial, ya que se proporcionó una perspectiva experta sobre cómo optimizar el enfoque. Tras consultar al ingeniero externo, se llegó a la conclusión de que la estrategia más eficiente sería implementar los algoritmos directamente sobre los datos originales de audio, aprovechando su estructura matricial, por lo que se descartó el uso de la transformada de Fourier y de los MFCCs. Esta recomendación permitió optimizar el proceso de procesamiento de datos y la aplicación de los algoritmos de manera más directa y efectiva.

En última instancia, dada la magnitud del conjunto de datos y la necesidad de administrar eficientemente la memoria y la capacidad de cómputo, se optó por implementar un input pipeline utilizando herramientas y pipelines de TensorFlow. Esta implementación permitió convertir los archivos de audio en tensores de valores decimales y organizarlos en lotes para su procesamiento por lotes. Esta modificación permitió continuar con el estudio del clasificador, ya que se estaban encontrando múltiples problemas debido a que la memoria de Google.Collab no era suficiente y se estaba imposibilitando el trabajo.

Esta conversión a tensores es fundamental ya que facilita el manejo de los datos en memoria y permite realizar operaciones matemáticas eficientes en ellos. Además, al utilizar tensores en lugar de datos de audio brutos, se redujo significativamente el consumo de memoria, lo que evitó errores de memoria y optimizó el rendimiento del sistema durante el procesamiento de los datos. El input pipeline también permitió organizar los datos en lotes, lo que facilitó su procesamiento por lotes durante el entrenamiento de modelos de aprendizaje automático. Esto aceleró el proceso de entrenamiento y mejoró la eficiencia del sistema en general.

4.1.3. Entrenamiento

En esta fase del KDD, se eligieron y entrenaron diferentes modelos matemáticos para el análisis del set de datos. El objetivo del proyecto es conseguir predecir las etiquetas de las señales de sonidos no verbales, para ello se entrenaron los siguientes modelos:

- K-vecinos más cercanos (KNN)
- Redes de neuronas artificiales secuenciales (SNN)

Ambos algoritmos son supervisados, para la clasificación de datos nuevos se entrena al modelo con datos etiquetados y el modelo aprende en función de si ha clasificado bien o mal datos etiquetados.

La idea básica detrás de KNN es que los puntos de datos similares tienden a agruparse en el espacio de características. Cuando se clasifica un nuevo punto de datos (en este caso, un nuevo audio), el algoritmo busca los "K"puntos de datos más cercanos (vecinos) en el espacio de características. Luego, asigna la clase más común entre esos vecinos al punto de datos de prueba. Para KNN se obtuvieron resultados poco satisfactorios con precisiones inferiores al 50 %. Aunque es posible que se pudieran obtener mejores resultados, utilizando representaciones de los datos como espectrogramas basados en los coeficientes cepstrales de Mel, se tomó la decisión de abordar el problema de clasificación principalmente con redes neuronales artificiales.

Las redes neuronales en el contexto del aprendizaje automático (machine learning) son un tipo de modelo computacional inspirado en la estructura y el funcionamiento del cerebro humano. Están compuestas por unidades de procesamiento llamadas neuronas artificiales que están organizadas en capas y conectadas mediante conexiones ponderadas. Las redes neuronales articificiales o redes neuronales secuenciales (en adelante RNA), son excelentes en la clasificación de sonidos por ser capaces de aprender características complejas de los datos, su flexibilidad en la representación de datos y adaptabilidad, esto las convierte en clasificadores consistentes aprueba de fallos como pueden ser incompletitudes en los datos.

Se comenzó realizando arquitecturas secuenciales con capas densas. Las capas densas son un tipo de conexiones entre neuronas que conectan todas las neuronas de entrada con todas las de salida, y de manera automática utilizando regresores lineales ajustan mediante pesos como de importante es cada conexión. Este tipo de capas tienen la desventaja de que estudian de manera parcial el orden en el que aparecen los datos, y los sonidos que se pretenden estudiar tienen carácteristicas que son temporalmente dependientes, como puede ser tras una tos un sonido de inspiración o la variación en amplitud al final de la misma. Se probaron modelos más simples con pocas capas y más complejos con muchas capas. Los resultados fueron peores que los obtenidos con KNN y además el modelo se encontraba limitado por su arquitectura, no superando el umbral de 25 % a pesar de aumentar el número de capas.

Viendo los resultados anteriores, se probó con modelos basados en capas convolucionales unidimensionales, esta arquitectura permite a la red extraer información posicional de los eventos que ocurren en los audios, este tipo de capas utilizan múltiples filtros sobre

ventanas de los vectores de amplitudes, obteniendo diferentes mapas de características. La finalidad de estas capas es reducir la dimensionalidad de los audios, llevando el espacio de dimensiones a otras más facilmente interpretables por regresores lineales. La mejora al usar este tipo de modelos fue notable, tanto modelos sencillos como más complejos, el mejor modelo alcanzado en mayo de 2024 tuvo un precisión del 85 %, a continuación se muestra la tabla parámetros entrenables de la mejor arquitectura:

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 63998, 64)	256
max_pooling1d (MaxPooling1D)	(None, 31999, 64)	0
conv1d_1 (Conv1D)	(None, 31997, 64)	12,352
max_pooling1d_1 (MaxPooling1D)	(None, 5332, 64)	0
conv1d_2 (Conv1D)	(None, 5330, 64)	12,352
max_pooling1d_2 (MaxPooling1D)	(None, 1776, 64)	0
conv1d_3 (Conv1D)	(None, 1774, 64)	12,352
max_pooling1d_3 (MaxPooling1D)	(None, 443, 64)	0
conv1d_4 (Conv1D)	(None, 441, 128)	24,704
max_pooling1d_4 (MaxPooling1D)	(None, 220, 128)	0
flatten (Flatten)	(None, 28160)	0
dense (Dense)	(None, 64)	1,802,304
dense_1 (Dense)	(None, 6)	390

Los hiperparámetros principales para este modelo:

Una vez se definió la arquitectura se estimaron los hiperparámetros para cada modelo, se utilizaron para todos los modelos de RNA los siguientes:

- Train(%), Test(%): 80 %, 20 %
- Batch size: 128 (máximo permitido por colab)
- Learning rate: Variable entre 0.01 y 0.001
- Epochs: dependiente del modelo

Para el mejor modelo el número de épocas fue 12.

4.1.4. Comparativa de modelos

Por simplicidad para la comparativa de modelos se tuvo en cuenta principalmente su precisión para resolver el problema en cuestión, se deja como trabajo a futuro el estudio en escalabilidad y ocupación en memoria de los parámetros entrenados.

Para la elección del modelo que se usará para la implementación real, se compararon

los siguientes resultados:

COMPARACIÓN DE MODELOS	
CLASIFICADORES	PRECISIÓN
KNN	41%
RN DENSA SENCILLA	20%
RN DENSA COMPLEJA	23%
RN CONV1D SIMPLE	63%
RN CONV1D COMPLEJA	83%

El mejor modelo obtenido en abril de 2024 fue el realizado con capas convolucionales 1D, que por tener más de 2 capas entra en la categoría de deep learning.

4.1.5. Implementación

Para la implementación del proyecto, se ha desarrollado una aplicación que ofrece una solución práctica para la detección en tiempo real de varios tipos de sonidos, como tos, aclaración de garganta, risas, suspiros, inspirar y estornudos.

La aplicación se inicia importando los módulos necesarios y configurando la detección de los tipos de sonidos, además de cargar un modelo previamente entrenado con una efectividad del 85 %. Luego, utiliza Matplotlib para mostrar las formas de onda de audio en tiempo real y capturar las muestras entrantes para su procesamiento. Mediante una función definida, la aplicación detecta los diferentes tipos de sonidos utilizando el modelo cargado y los presenta en un gráfico para una fácil comprensión.

El flujo de audio se inicia mediante SoundDevice y se crea una animación continua para visualizar de manera fluida las muestras de audio y sus clasificaciones. Esta implementación puede ser muy útil en diversos ámbitos profesionales, como en las llamadas telefónicas (para evaluar el nivel de satisfacción del cliente durante la llamada) y en servicios sanitarios (para supervisar el bienestar del paciente).

5. Conclusión

Tras realizar numerosas pruebas con distintas configuraciones, se ha logrado obtener un modelo con una reseñable precisión del 83 %. Un modelo de estas características es suficientemente robusto y fiable como para ser implementado con éxito en diversos entornos reales, por lo que el objetivo inicial ha sido alcanzado con éxito. Cabe destacar que este clasificador solo representa la punta del iceberg del campo de la clasificación de sonidos; siendo un ámbito que aún aguarda infinitas posibilidades por explorar. Existen múltiples direcciones para futuras investigaciones y mejoras, tales como la optimización de algoritmos, la integración de técnicas de aprendizaje profundo más avanzadas, y

la ampliación del conjunto de datos para incluir una mayor variedad de sonidos y contextos.

Además, a pesar de haber puesto el foco en entornos de cuidado, la aplicación de estos modelos no se limita solo a ese ámbito; pudiendo extenderse a otras áreas como la seguridad, la domótica, la asistencia personal y la industria del entretenimiento, entre muchas otras. La intersección de la clasificación de sonidos con la inteligencia artificial promete revolucionar diversos sectores, facilitando una interacción más natural y eficiente entre humanos y máquinas.

En conclusión, aunque se han alcanzado los objetivos deseados con este proyecto, continuar explorando y perfeccionando esta tecnología mejorará enormemente su precisión y aplicabilidad, abriendo nuevas fronteras de aplicación en este campo tan prometedor.

Referencias

[1] Fabián Aguirre Martín et al. «Desarrollo y análisis de clasificadores de señales de audio». En: (2017).

[2] Rachel E Bouserhal et al. «Classification of nonverbal human produced audio events: a pilot study». En: (2018).

[3] Sachin Chachada y C-C Jay Kuo. «Environmental sound recognition: A survey». En: *APSIPA Transactions on Signal and Information Processing* 3 (2014), e14.

[4] Mercedes Gomez Martin et al. «CLASIFICACIÓN DE VOCES A TRAVÉS DE SERIES DE FOURIER Y REDES NEURONALES». En: (2023).

[5] Patricio Rodríguez Ramírez. «Clasificación automática de sonidos utilizando aprendizaje máquina». En: (2020).

[6] Cristóbal González Rodrıguez, Miguel A Alonso Arévalo y Eloısa Garcıa Canseco. «Eliminación de ruido en sonidos cardıacos mediante técnicas de aprendizaje profundo». En: ().

[7] Aurora Salgado Díaz del Río. «Reconocimiento automático de instrumentos mediante aprendizaje máquina». En: (2019).

[8] Maximilian Schmitt et al. «A bag-of-audio-words approach for snore sounds' excitation localisation». En: *Speech Communication; 12. ITG Symposium*. VDE. 2016, págs. 1-5.