Audio Classification Model (ACM)

Andres Bravo Medina

October 2022

Contenido

Introducción

El análisis de audio es un proceso de transformación, exploración e interpretación de señales de audio capturadas por dispositivos digitales, que pueden venir en distintos formatos como wav, mp3, aiff, flac, entre otras.
Aplicaciones del analisis de audio

- Reconocimiento de música
- Identificador de sonidos de animales
- Reconocimiento de sonidos ambientales
- Reconocimiento de voz

Introducción

Podemos empezar obteniendo el resumen de metadatos de audio:

- Canales
- Ancho de muestra
- Tasa de fotogramas
- Ancho de fotograma:
- Intensidad
- Duracion
- Recuento de fotogramas

Transformada de Fourier

La transformada de Fourier es una de las operaciones más fundamentales en el procesamiento de señales.

Transforma nuestra señal en el dominio del tiempo al dominio de la frecuencia. Mientras que el dominio del tiempo expresa nuestra señal como una secuencia de muestras, el dominio de la frecuencia expresa nuestra señal como una superposición de sinusoides de diferentes magnitudes, frecuencias y desfases.

Teoría

Una representación transformada en el dominio de la frecuencia $\{X_n\} := X_0, X_1, \ldots, X_{n-1}$, se puede encontrar usando la siguiente formulación:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{\frac{-2\pi i}{N} kn}$$

$$X_k = \sum_{n=0}^{N-1} xn \left(cos\left(\frac{2\pi}{N} kn\right) - i \cdot sin\left(\frac{2\pi}{N} kn\right)\right)$$

Teoría

Una transformada discreta de Fourier es computacionalmente bastante difícil de calcular, con una complejidad de tiempo del orden $O(n^2)$. Pero hay un algoritmo más rápido llamado Transformada rápida de Fourier que funciona con una complejidad de $O(nlog(n))$.

Transformada de Fourier de tiempo corto

Con las transformadas de Fourier, convertimos una señal del dominio del tiempo al dominio de la frecuencia. Al hacerlo, vemos cómo cada punto en el tiempo interactúa entre sí para cada frecuencia. Las transformadas de Fourier de tiempo corto lo hacen para los puntos vecinos en el tiempo en lugar de para toda la señal. Esto se hace utilizando una función de ventana que salta con una longitud de salto específica para darnos los valores del dominio de frecuencia.

Teoría

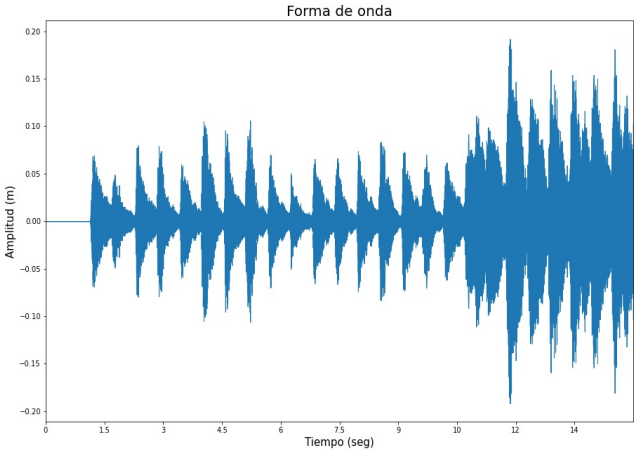
Sea x una señal de tamano L y w una funcion de ventana de tamaño N. El indice de fotograma maximo M seria $\frac{L-N}{N}$. $X(m,k)$ denotaria el k-esimo coeficiente de Fourier en el tiempo m. Otro parametro definido es H, llamado tamaño de salto. Determina el tamaño de paso de la función de ventana.
Entonces $X(m,k)$ es dado por:

$$X(m,k) = \sum_{n=0}^{N-1} x(n+mH)w(n)e^{\frac{-2\pi i}{N}kn}$$

Representaciones visuales

- **Forma de onda**
  Refleja cómo cambia la amplitud en el tiempo. Estas amplitudes no son muy informativas, ya que solo describen el volumen de la grabación de audio y no nos dice qué pasa con las frecuencias.

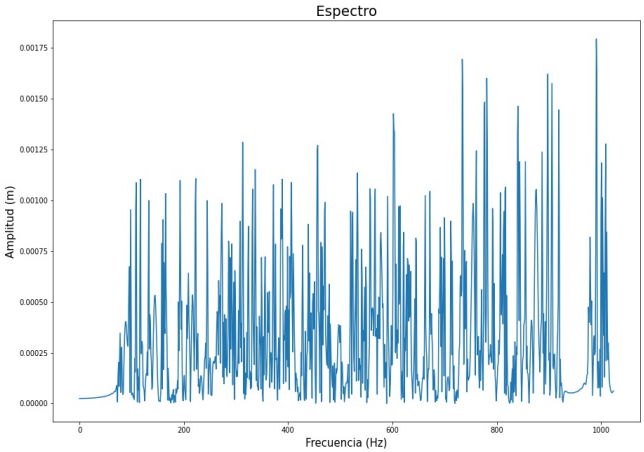
Representaciones visuales



Representaciones visuales

- **Espectro**
  Este tipo de visualización nos dice qué diferentes frecuencias están presentes en la señal, pero pasa por alto el tiempo.

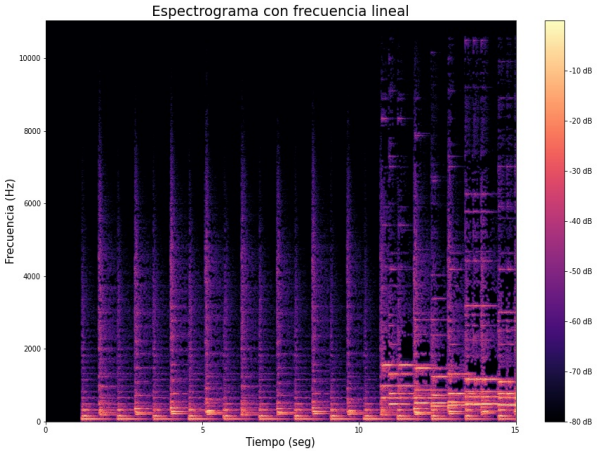
Representaciones visuales



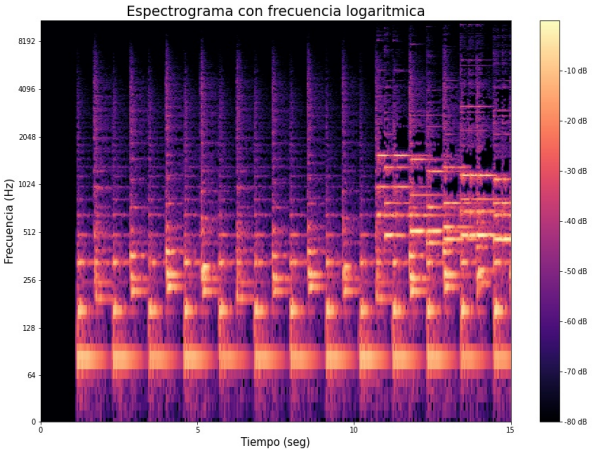
Representaciones visuales

- **Espectrograma**
  El espectrograma permite identificar las diferentes variaciones de la frecuencia y la intensidad del sonido a lo largo de un periodo de tiempo.

Representaciones visuales



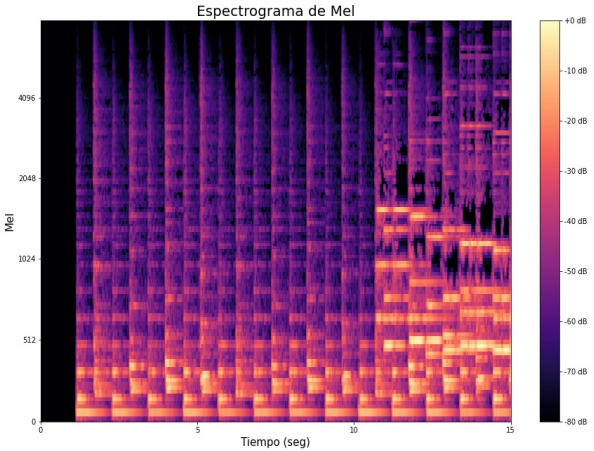
Representaciones visuales



- **Espectrograma de Mel**
  Un espectrograma de Mel es un espectrograma en el que las frecuencias se convierten a la escala de Mel.

$$f_{Mel} = 2595 \cdot log_{10}\left(1 + \frac{f}{700}\right)$$

Representaciones visuales

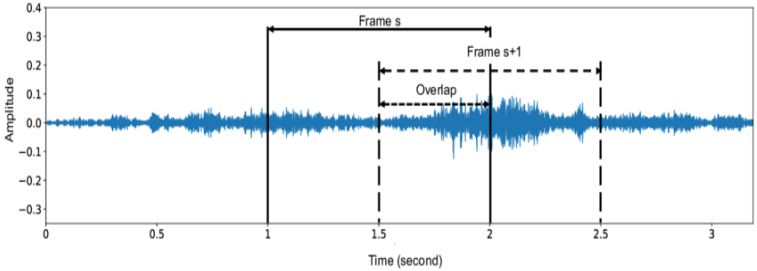


Preprocesamiento

- **Framming**
  Consiste en dividir la señal original en $\#F$ fotogramas con longitud N_f, una superposicion M y un framing hop $H = N_f - M$. La superposición de fotogramas ayuda a evitar la pérdida de información entre fotogramas adyacentes.

$$S = \sum_{n=0}^{N_s} x[n] = \sum_{i=0}^{\#F-1} F[i]$$

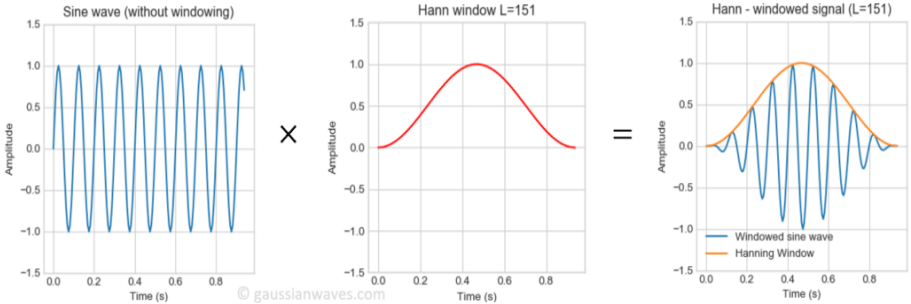
Preprocesamiento



Preprocesamiento

- **Windowing**
  Es una técnica para minimizar las fugas espectrales.
  Básicamente, todas las ventanas hacen lo mismo: reducir o suavizar la amplitud al principio y al final de cada fotograma mientras la aumentan en el centro para conservar el valor medio.

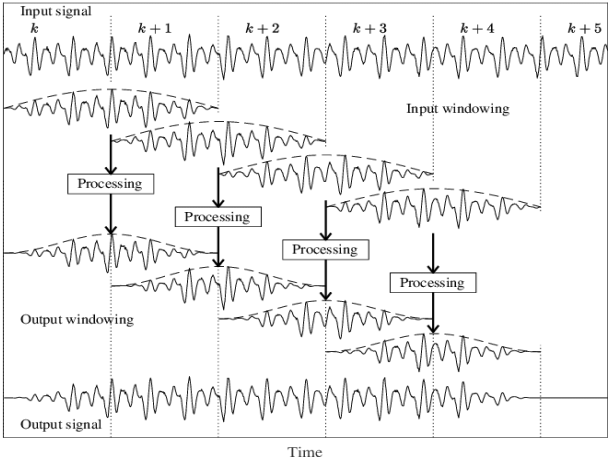
Preprocesamiento



Preprocesamiento

- **Overlap-add**
  Evita que se pierda la información vital que puede provocar la creación de ventanas. Proporciona un solapamiento del 30-50 por ciento entre fotogramas adyacentes, lo que permite modificarlos sin riesgo de distorsión.

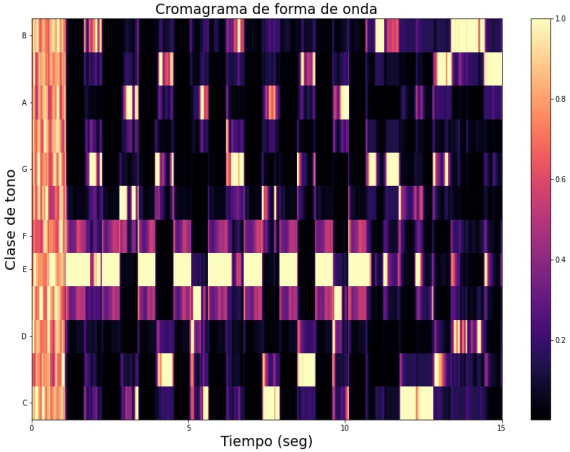
Preprocesamiento



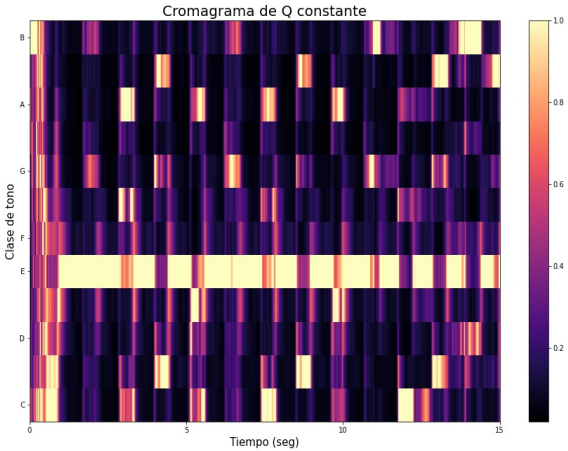
Extracción de características

- **Cromagrama**
  Es un descriptor que representa el contenido tonal de una señal de audio de forma condensada.
  Un vector de croma es un vector de características de 12 elementos que indica la energía de cada clase de tono, $\{C, C\#, D, D\#, E, ..., B\}$, que está presente en la señal.
  Hay tres tipos de cromagramas:
  - Cromagrama de forma de onda
  - Cromagrama de Q constante
  - Cromagrama de estadísticas normalizadas de energía cromática (CENS)

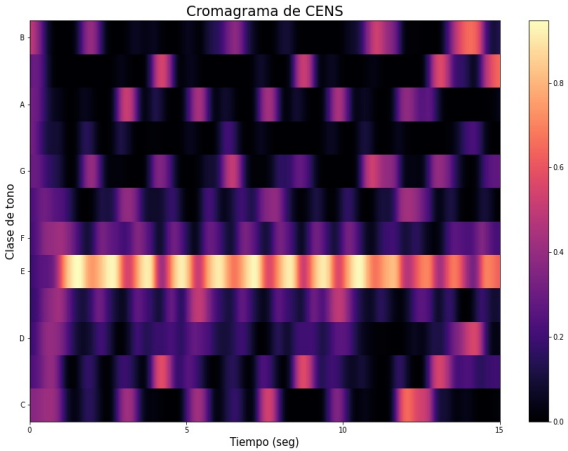
Extracción de características



Extracción de características



Extracción de características



Extracción de características

- **Mel-Frequency Cepstral Coefficients (MFCCs)**
  Los MFCC son una representación compacta del espectro de una señal de audio.
  Los coeficientes MFCC contienen información sobre los cambios de velocidad en las diferentes bandas del espectro.
  Si un coeficiente cepstral tiene un valor positivo, la mayor parte de la energía espectral se concentra en las regiones de baja frecuencia y viceversa.

Extracción de características

- Pre-emphasis
  El pre-énfasis se refiere al filtrado que enfatiza las frecuencias más altas. Su objetivo es equilibrar el espectro de los sonidos de voz que tienen una caída pronunciada en la región de las frecuencias altas.
  Por lo tanto, el pre-énfasis elimina algunos de los efectos glotales de los parámetros del tracto vocal.

Extracción de características

- Frame blocking and windowing
  Para que las características acústicas sean estables, el habla debe examinarse durante un periodo de tiempo suficientemente corto.

Extracción de características

- DFT Spectrum
  Cada cuadro con ventana se convierte en espectro de magnitud aplicando la DFT.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-i2\pi nk}{N}} \; ; \; 0 \leq k \leq N-1$$

  Donde N es el número de puntos utilizados para calcular la DFT.

Extracción de características

- Mel Spectrum
  El espectro de Mel se calcula haciendo pasar la señal transformada de Fourier a través de un conjunto de filtros de paso de banda conocidos como banco de filtros de Mel.

Extracción de características

Las frecuencias centrales de los filtros suelen estar espaciadas uniformemente en el eje de la frecuencia. El espectro de Mel del espectro de magnitud X(k) se calcula multiplicando el espectro de magnitud por cada uno de los filtros triangulares de ponderación de Mel.

$$s(m) = \sum_{k=0}^{N-1} (|X(k)|^2 H_m(k)) \; ; \; 0 \leq m \leq M-1$$

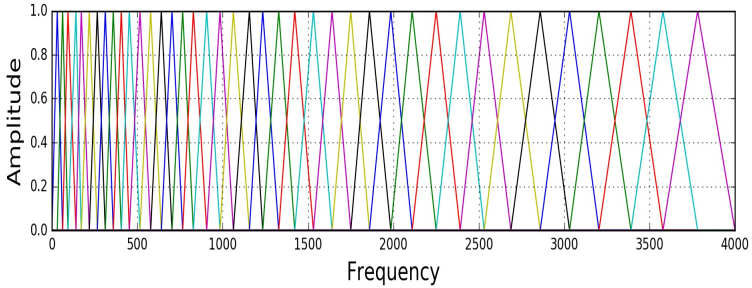
Extracción de características

donde M es el número total de filtros triangulares de ponderación de Mel. $H_m(k)$ es el peso dado a la k^{th} bin del espectro de energía que contribuye a la m^{th} banda de salida y se expresa como

$$H_m(k) = \begin{cases} 0 & , \quad k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)} & , \quad f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)} & , \quad f(m) < k \leq f(m+1) \\ 0 & , \quad k > f(m+1) \end{cases}$$

Con m que va desde 0 hasta $M-1$

Extracción de características



Extracción de características

- Discrete cosine transform (DCT)
  Los niveles de energía en bandas adyacentes tienden a estar correlacionadas. Al aplicar la DCT a los coeficientes de frecuencia de Mel transformados se obtiene un conjunto de coeficientes cepstrales. Antes de calcular la DCT, el espectro de Mel suele representarse en una escala logarítmica. Finalmente, la MFCC se calcula como

$$c(n) = \sum_{m=0}^{M-1} log_{10}(s(m)) cos\left(\frac{\pi n(m-0{,}5)}{M}\right) \;\; ; \;\; n = 0, 1, 2, \ldots, C-1$$

  donde $c(n)$ son los coeficientes cepstrales, y C es el número de MFCC.

Extracción de características

- Dynamic MFCC features
  La información adicional sobre la dinámica temporal de la señal se obtiene calculando derivadas de primer y segundo orden de los coeficientes cepstrales. La definición comúnmente utilizada para calcular el parámetro dinámico es:

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|}$$

Extracción de características

- **Zero-crossing Rate (ZCR)**
  La tasa de cruce por cero es la tasa de cambios de signo a lo largo de una señal, es decir, la tasa a la que la señal cambia de positiva a negativa o viceversa.

$$ZCR = \frac{1}{2L} \sum_{n=1}^{L} sgn[x(n)] - sgn[x(n-1)]$$

Donde

$$sgn[x(n)] = \begin{cases} 1 & , \quad x(n) \geq 0 \\ -1 & , \quad x(n) < 0 \end{cases}$$

$x(n)$ es la n-esima muestra de una ventana de longitud L

Extracción de características

- **Root-mean-square (RMS)**
  La envolvente RMS consiste en la raíz cuadrada media de la amplitud, calculada dentro de la ventana de agregación.

$$RMS = \sqrt{\frac{1}{L}\sum_{n=1}^{L} x(n)^2}$$

  Donde $x(n)$ es la n-esima muestra de una ventana de tamaño L.

- **Spectral features**
- Spectral Centroid
  El centroide espectral es una característica tímbrica que se utiliza para describir el brillo de un sonido. Representa el centro de gravedad de los componentes de frecuencia de una señal.

$$SC = \frac{\sum_{k=1}^{N} |X(k)| \cdot f_k}{\sum_{k=1}^{N} |X(k)|}$$

Donde $X(k)$ es el resultado de la STFT para el intervalo de frecuencia k-ésimo.

Extracción de características

- Spectral Flatness
  La planitud espectral da una estimación del ruido/sinusoidalidad de una señal de audio. Se puede utilizar para determinar las partes sonoras/sordas de una señal.

$$SFM = 10log_{10}\left(\frac{\left(\prod_{k=1}^{N}|X(k)|\right)^{\frac{1}{N}}}{\frac{1}{N}\sum_{k=1}^{N}|X(k)|}\right)$$

  Donde $X(k)$ es el resultado de la STFT para el intervalo de frecuencia k-ésimo.

Extracción de características

- Spectral Flux
  El flujo espectral mide la cantidad de cambio entre cuadros espectrales sucesivos.

$$SF = \sum_{k=1}^{N} H(X(t,k) - X(t-1,k))$$

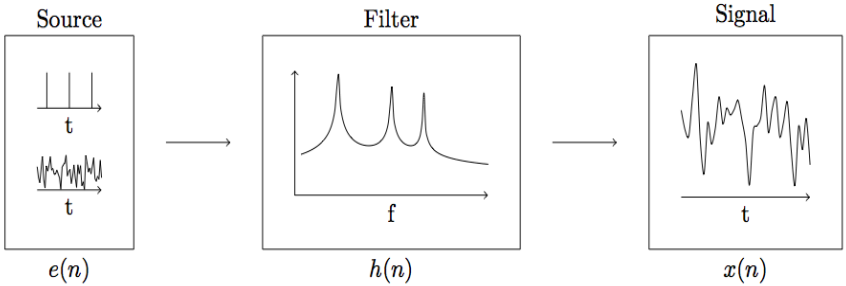
Donde

$$H(x) = \frac{x + x}{2}$$

es la función del rectificador de media onda, y t es el índice temporal del cuadro.

Extracción de características

- **Linear Predictive Coding (LPC)**
  LPC es un modelo de fuente-filtro en el que hay una fuente de sonido que pasa por un filtro.
  La fuente $e(n)$, modela las cuerdas vocales, mientras que el filtro resonante $h(n)$, modela el tracto vocal. La señal resultante es $x(n) = h(n) \cdot e(n)$



Extracción de características

- **Line spectral pairs (LSP) or line spectral frequencies (LSF)**
  Se utilizan para representar coeficientes de predicción lineal (LPC) para la transmisión a través de un canal.
  En el análisis LP del habla, se supone que un segmento corto que se genera como la salida de un filtro multipolar $H(z) = 1/A(z)$, donde $A(z)$ es el filtro inverso dado por

$$A(z) = 1 - \sum_{i=1}^{M} a_i \cdot z^{-k}$$

Extracción de características

Aquí M es el orden de análisis del LP y $\{a_i\}$ son los coeficientes de LP. Para definir los LSF, el polinomio del filtro inverso se descompone en dos polinomios

$$P(z) = A(z) + z^{-(M+1)}A(z^{-1})$$
$$Q(z) = A(z) - z^{-(M+1)}A(z^{-1})$$

Las raíces de los polinomios $P(z)$ y $Q(z)$ se llaman LSFs.
Todos los ceros de $P(z)$ y $Q(z)$ se encuentran en el círculo unitario.
Los ceros de P(z) y Q(z) están entrelazados entre sí.

Modelos

- Convolutional Neural Networks (CNN)
- Long Short-Term Memory (LSTM)
- Gaussian Mixture Models (GMM)
- K-Nearest Neighbors (KNN)
- Naïve Bayes (NB)
- Random Forest (RF)
- Support Vector Machine (SVM)

Modelos

Caracteristicas	Modelos
MFCC	CNN, LSTM, GMM, KNN, RF, SVM, NB
Espectrograma	
Espectrograma de Mel	
Cromagrama	
Caracteristicas Espectrales	GMM, KNN, RF, SVM, NB
LPC	
LSF	
RMS	
ZCR	

Librerías

- **Librosa**
  Es un paquete de python para análisis de música y audio. Proporciona los componentes básicos necesarios para crear sistemas de recuperación de información musical.
  Tiene:
  - Carga de audio
  - Representaciones espectrales
  - Escala de magnitud
  - Conversiones de unidades de tiempo y frecuencia
  - Visualizacion de datos
  - Extraccion de caracteristicas
  - Banco de filtros
  - Funciones de ventana

  Ultima versión: 27/06/2022

Librerias

- **Tensorflow-io**
  TensorFlow I/O es un paquete de extensión de Tensorflow, que incluye soporte io para una colección de sistemas de archivos y formatos de archivo que no están disponibles en el soporte integrado de TensorFlow.
  Ultima versión: 08/09/2022

Librerias

- **Iracema**
  Es un paquete de Python destinado a la investigación empírica de la interpretación musical, con foco en el análisis de la expresividad y la individualidad de las grabaciones de audio. Contiene modelos computacionales de extracción de información musical.
  Tiene:
  - Carga de audio
  - Extraccion de informacion espectral
  - Extraccion de caracteristicas
  - Ploteo de datos

  Ultima versión: 06/05/2021

Librerias

- **Pydub**
  Es un paquete de python que se enfoca en la manipulacion de audio.
  Ultima versión: 09/03/2021

Referencias

- Khan, T. (2019, 4 septiembre). A Deep Learning Model for Snoring Detection and Vibration Notification Using a Smart Wearable Gadget. Electronics, 8(9), 987. https://doi.org/10.3390/electronics8090987
- D. F. Silva, V. M. A. D. Souza, G. E. A. P. A. Batista, E. Keogh and D. P. W. Ellis, Applying Machine Learning and Audio Analysis Techniques to Insect Recognition in Intelligent Traps, 2013 12th International Conference on Machine Learning and Applications, 2013, pp. 99-104, doi: 10.1109/ICMLA.2013.24.

Referencias

- Benba, A., Jilbab, A. Hammouch, A. (2015, 30 junio). Detecting Patients with Parkin son's disease using Mel Frequency Cepstral Coefficients and Support Vector Machines. International Journal on Electrical Engineering and Informatics, 7(2), 297-307. https://doi.org/10.15676/ijeei.2015.7.2.10
- Editor. (2022, 12 mayo). Audio Analysis With Machine Learning: Building AI-Fueled Sound Detection App. AltexSoft. Recuperado 18 de octubre de 2022, de https://www.altexsoft.com/blog/audio-analysis/
- Tanuwidjaja, O. (2022, 22 enero). Get To Know Audio Feature Extraction in Python - Towards Data Science. Medium. Recuperado 18 de octubre de 2022, de https://towardsdatascience.com/get-to-know-audio-feature-extraction-in-python a499fdaefe42