TÉCNICAS AVANZADAS DE DATA MINING Y SISTEMAS INTELIGENTES

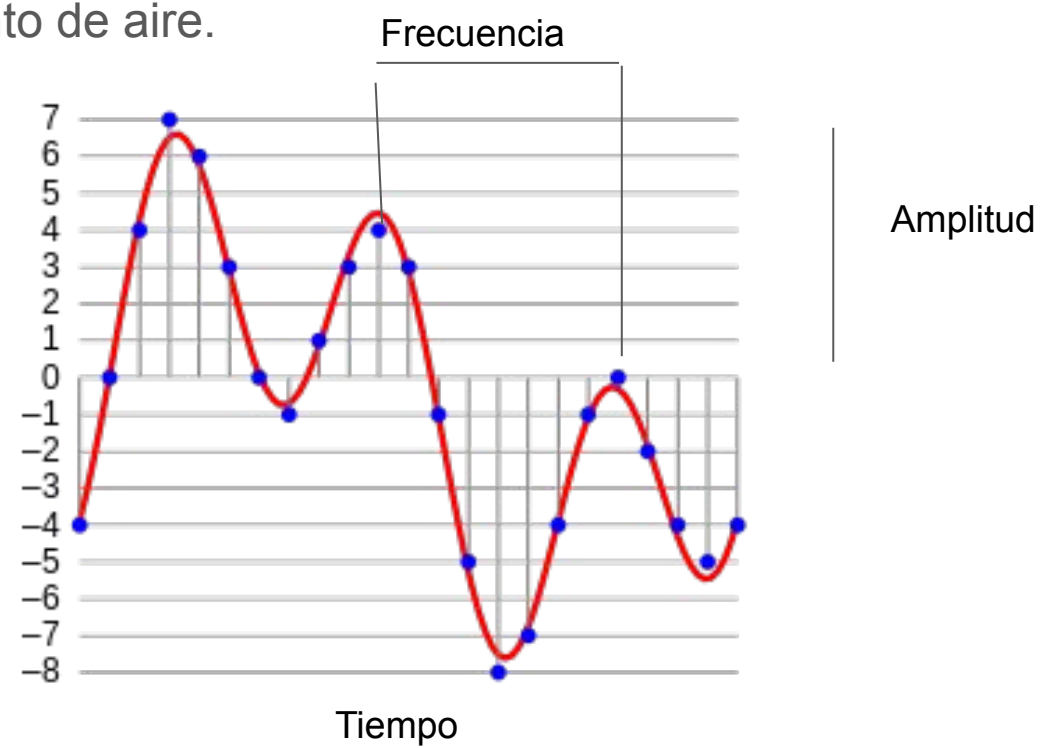
Integrantes:

- Mendoza Jacinto, Juan Manuel
- Morales Pariona, Jose Ulises
- Tippe Quintanilla, Percy Kim

Marco Teórico

Audio

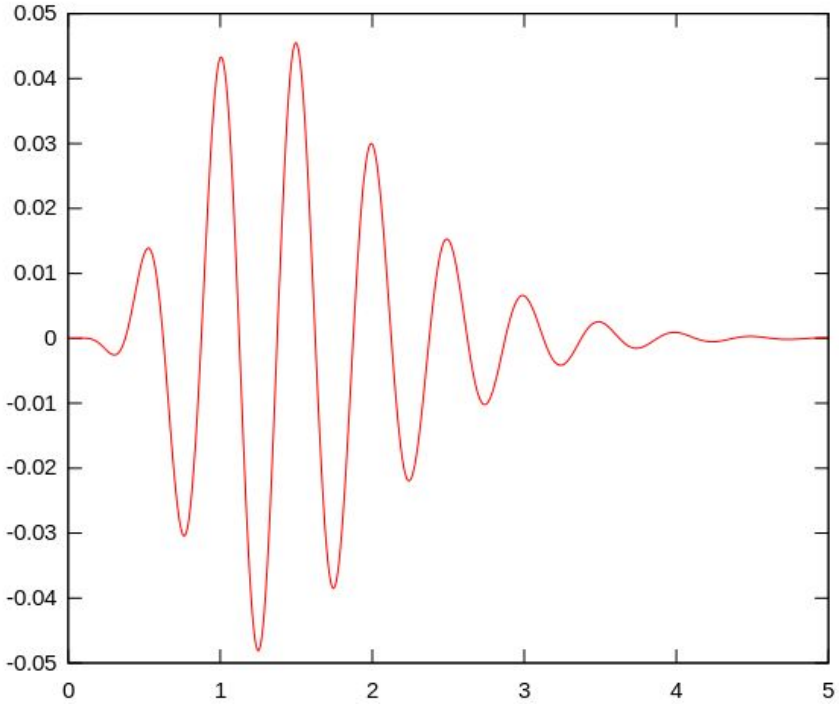
El sonido son compresiones y rarefacciones en el aire que un oído captará.El sonido es un movimiento de aire.



https://www.analyticsvidhya.com/blog/2017/08/audio-voice-processing-deep-learning/

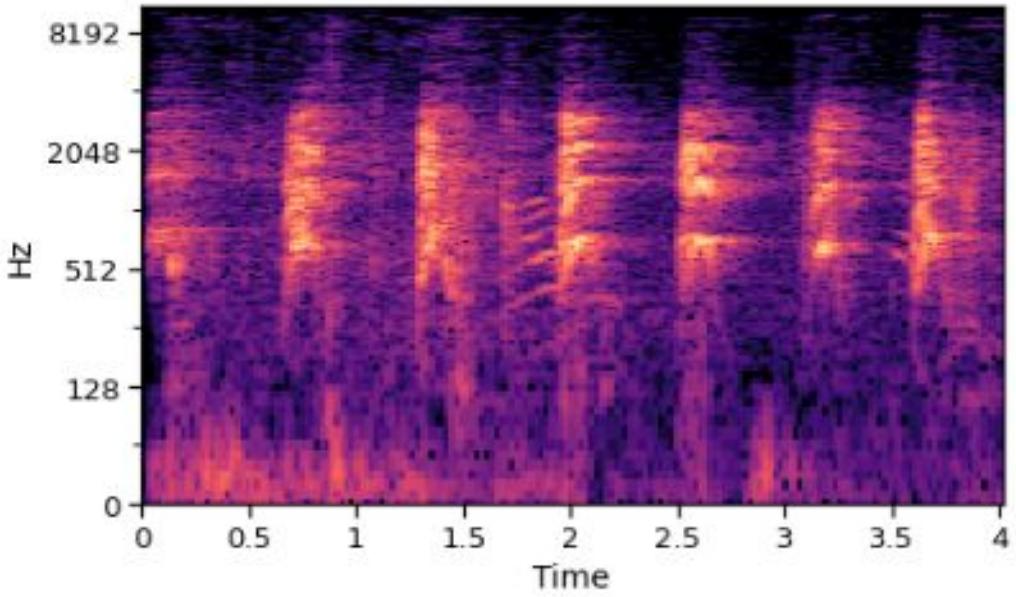
Filtro Gammatone

Un filtro de gammatona es un filtro lineal descrito por una respuesta de impulso que es el producto de una distribución gamma y un tono sinusoidal.



Espectrograma

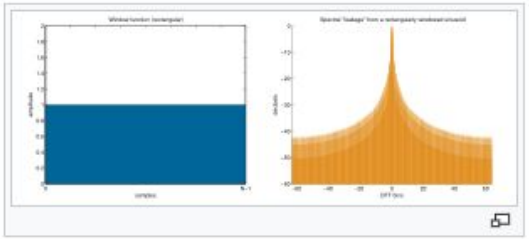
Un espectrograma es una representación visual del espectro de frecuencias de una señal, ya que varía con el tiempo. Cuando se aplica a una señal de audio, los espectrogramas a veces se llaman ecografías, huellas de voz o diagramas de voz.



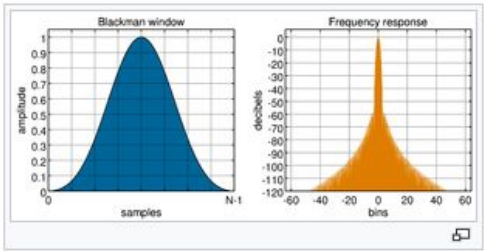
Ventana (función)

Las ventanas son funciones matemáticas usadas con frecuencia en el análisis y el procesamiento de señales para evitar las discontinuidades al principio y al final de los bloques analizados.

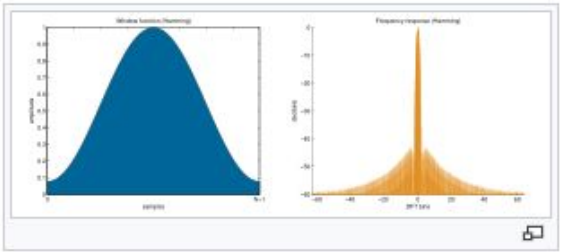
Rectangular [editar]



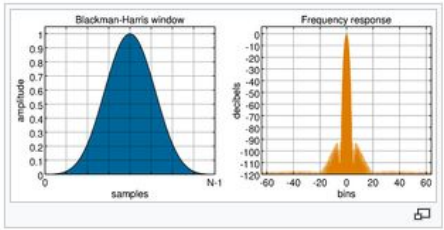
Blackman [editar]



Hamming [editar]

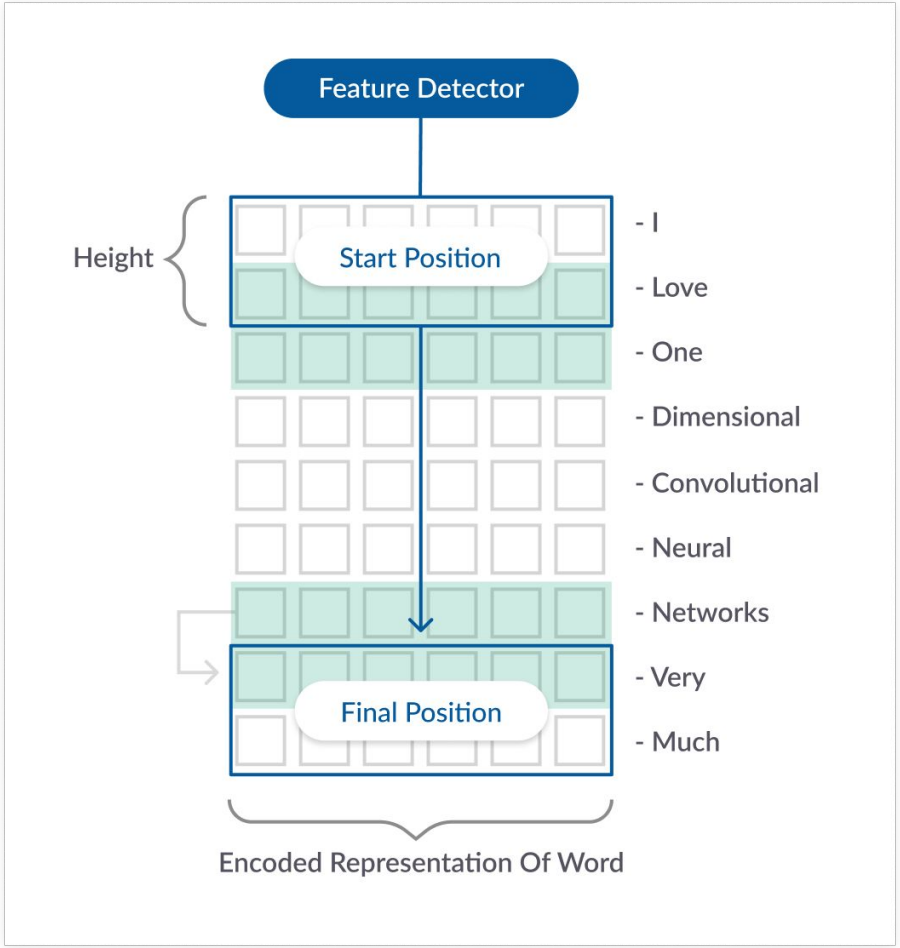


Blackman-Harris [editar]



1D CNN

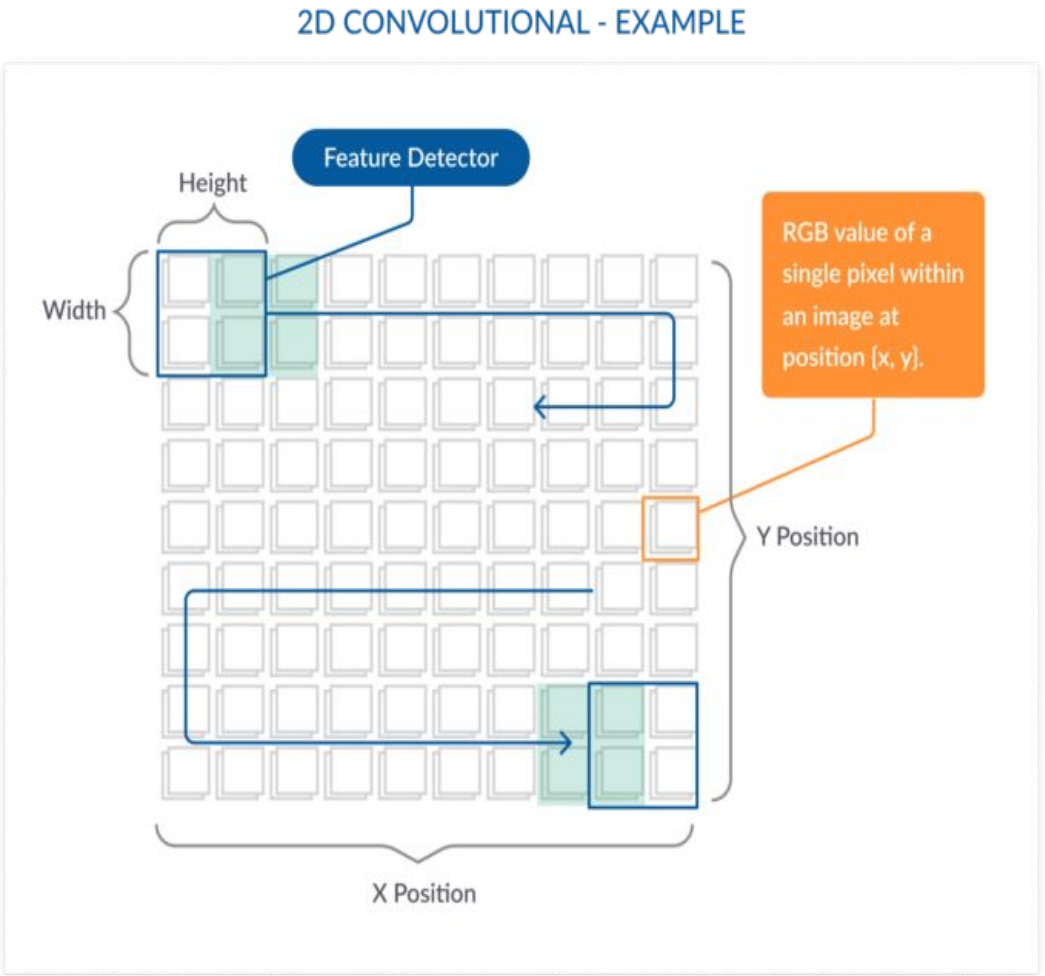
Procesamiento del lenguaje natural (PNL), una oración se compone de 9 palabras. Cada palabra es un vector que representa una palabra. El filtro cubre al menos una palabra; un parámetro de altura especifica cuántas palabras debe considerar el filtro a la vez. En este ejemplo, la altura es 2, lo que significa que el filtro se mueve 8 veces para escanear completamente los datos.



https://missinglink.ai/guides/keras/keras-conv1d-working-1d-convolutional-neural-networks-keras

2D CNN

En una red convolucional 2D, cada píxel dentro de la imagen está representado por su posición xey, así como por la profundidad, que representa los canales de imagen (rojo, verde y azul). El filtro en este ejemplo es de 2 × 2 píxeles. Se mueve sobre las imágenes tanto horizontal como verticalmente.

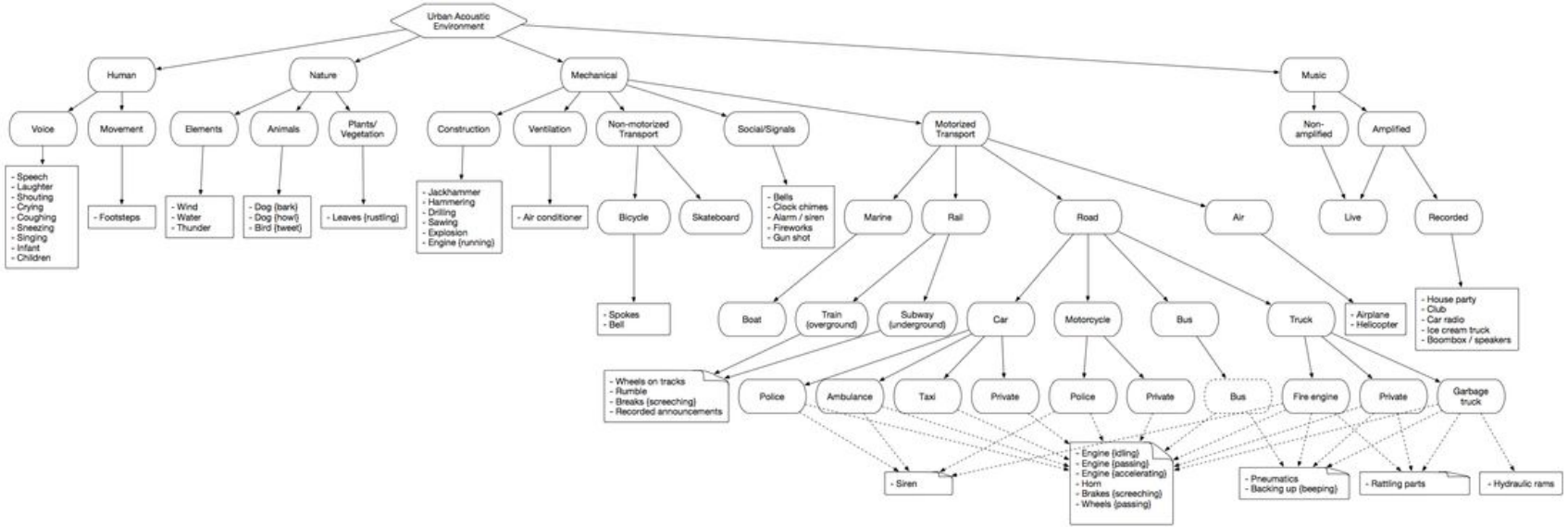


https://missinglink.ai/guides/keras/keras-conv1d-working-1d-convolutional-neural-networks-keras

DATASET

UrbanSound8K

El UrbanSound8K es un dataset que contiene 8732 sonidos en formato WAV y están etiquetados como uno de 10 clases posibles. Las clases corresponden a la taxonomía de sonido urbano (Urban Sound Taxonomy).



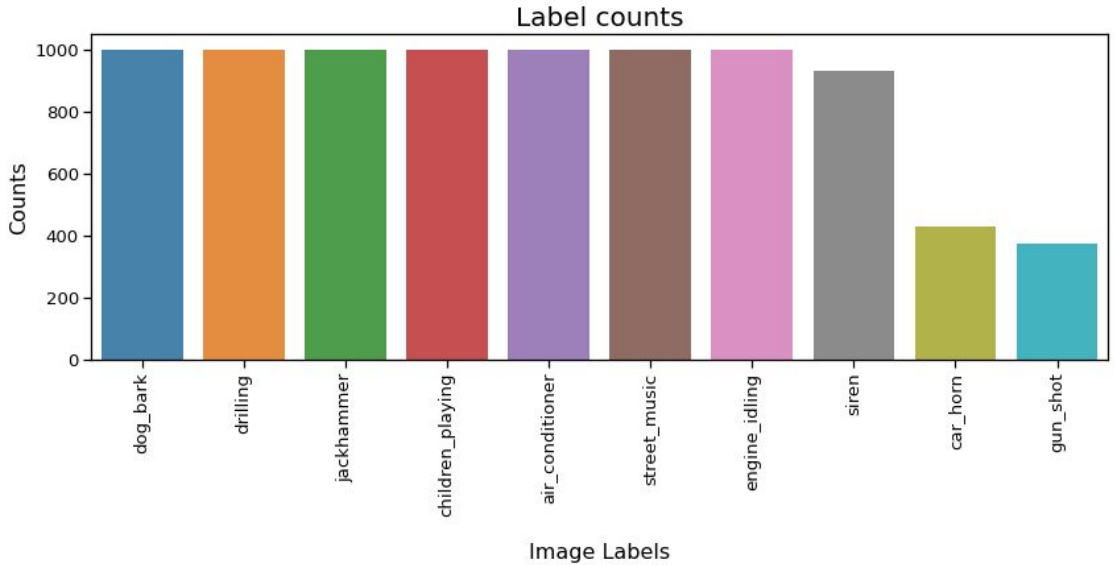
UrbanSound8K

Las 10 clases del dataset son las siguientes:

Aire Acondicionado (Air conditioner)	Motor en ralentí (Engine idling)
Claxon (Car horn)	Disparo (Gun shot)
Niños jugando (Children playing)	Martillo Neumático (Jackhammer)
Ladrido (Dog bark)	Sirena (Siren)
Taladro (Drilling)	Música de calle (Street music)

UrbanSound8K

La distribución de las 10 clases no está completamente balanceada debido a que dos clases, “claxon” (car horn) y “disparo” (gun shot), tienen menor cantidad de sonidos respecto al resto de clases.



ARQUITECTURAS

Arquitecturas para Clasificación de Audio

En un trabajo de investigación publicado en 2017 se realizó una comparación entre diferentes arquitecturas de redes convolucionales para la clasificación de audio AED (Acoustic Event Detection). Este trabajo compara las siguientes arquitecturas:

- Fully-Connected Deep Neural Network (DNN)
- AlexNet
- VGG
- Inception V3
- ResNet-50.

S. Hershey et al., "CNN architectures for large-scale audio classification," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 131-135, doi: 10.1109/ICASSP.2017.7952132.

Arquitecturas para Clasificación de Audio

La métrica utilizada para la validación es la AUC (Area under the Receiver Operating Characteristic Curve) la cual mide la probabilidad de clasificar correctamente un evento positivo. Adicionalmente se utilizó la métrica mean Average Precision (mAP).

Architectures	Steps	Time	AUC	d-prime	mAP
Fully Connected	5M	35h	0.851	1.471	0.058
AlexNet	5M	82h	0.894	1.764	0.115
VGG	5M	184h	0.911	1.909	0.161
Inception V3	5M	137h	**0.918**	**1.969**	0.181
ResNet-50	5M	119h	0.916	1.952	**0.182**
ResNet-50	17M	356h	**0.926**	**2.041**	**0.212**

Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification

J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," in IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279-283, March 2017, doi: 10.1109/LSP.2017.2657381.

Preprocesamiento de los dados

Data Augmentation

Para extender el tamaño del dataset se utilizaron 4 técnicas de Data Augmentation generando 5 sets. Los métodos de data augmentation utilizados son:

- **Time Stretching (TS):** [0.81, 0.93, 1.07, 1.23]
- **Pitch Shifting (PS1):** [-2, -1, 1, 2]
- **Pitch Shifting (PS2):** [-3.5, -2.5, 2.5, 3.5]
- **Dynamic Range Compression (DRC):** [music standard, film standard, speech, radio]
- **Background Noise (BG):** [street-workers, street-traffic, street-people, park]

Preprocesamiento de los dados

El preprocesamiento de audio se hizo utilizando el aplicativo Essentia (Python) para extraer espectrogramas (mel-spectrograms) de 128 bandas (0-22050 Hz) desde los archivos de audio.

Dado que los archivos de audio tienen una duración variable y menor a 4 segundos, se fijo la duración en 3s, obteniendo 128 frames de audio.

Modelo

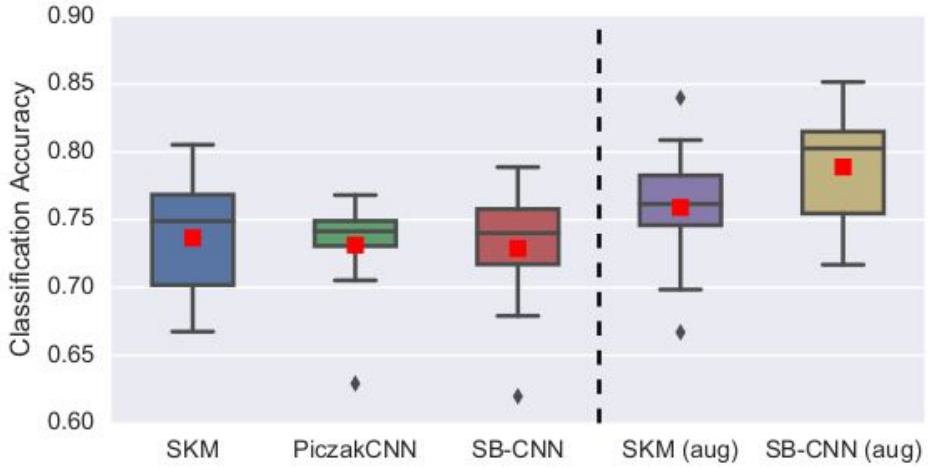
El modelo utilizado es una red neuronal convolucional (CNN) integrada por 3 capas convolucionales alternadas con 2 capas de pooling y seguidas por 2 capas Fully-Connected (dense):

- **Capa 1:** Conv(24, 1, 5, 5) - Max-pooling (4, 2) - ReLU
- **Capa 2:** Conv(48, 24, 5, 5) - Max-pooling (4, 2) - ReLU
- **Capa 3:** Conv(48, 48, 5, 5) - ReLU
- **Capa 4:** Dense(2400, 64) - ReLU
- **Capa 5:** Dense(64, 10) - Softmax

Evaluación

La métrica utilizada es el Accuracy de la clasificación. Los resultados se compararon con los de arquitecturas utilizadas en investigaciones previas. Adicionalmente se compararon los modelos utilizando los datasets con y sin Data Augmentation.

- Spherical K-Means (SKM)
- PiczakCNN
- **SB-CNN (propuesta)**

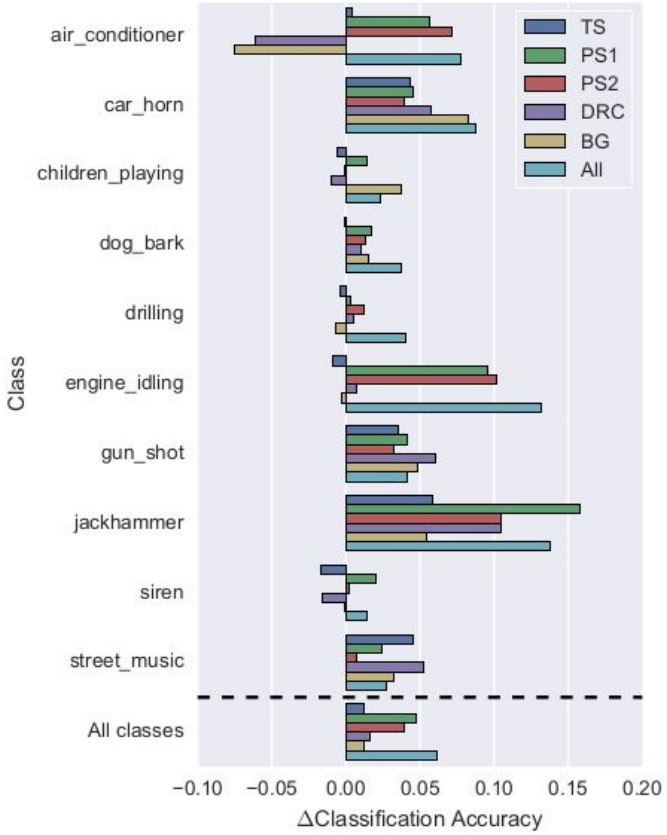


Conclusiones

Se concluye del artículo que las técnicas de Data Augmentation tienen un impacto positivo en el Accuracy general del modelo.

Sin embargo cuando se analiza el impacto por clase se observa que este no siempre es beneficioso.

En el gráfico de la derecha se observa que si bien la variación del Accuracy es positiva en la mayoría de casos, para la clase “Aire Acondicionado” este se ve perjudicado por el “Dynamic Range Compression” (DRC) y por el ruido de fondo “Background Noise” (BG).



End-to-End Clasificación de sonido ambiental utilizando una red neuronal convolucional 1D

1. Introducción

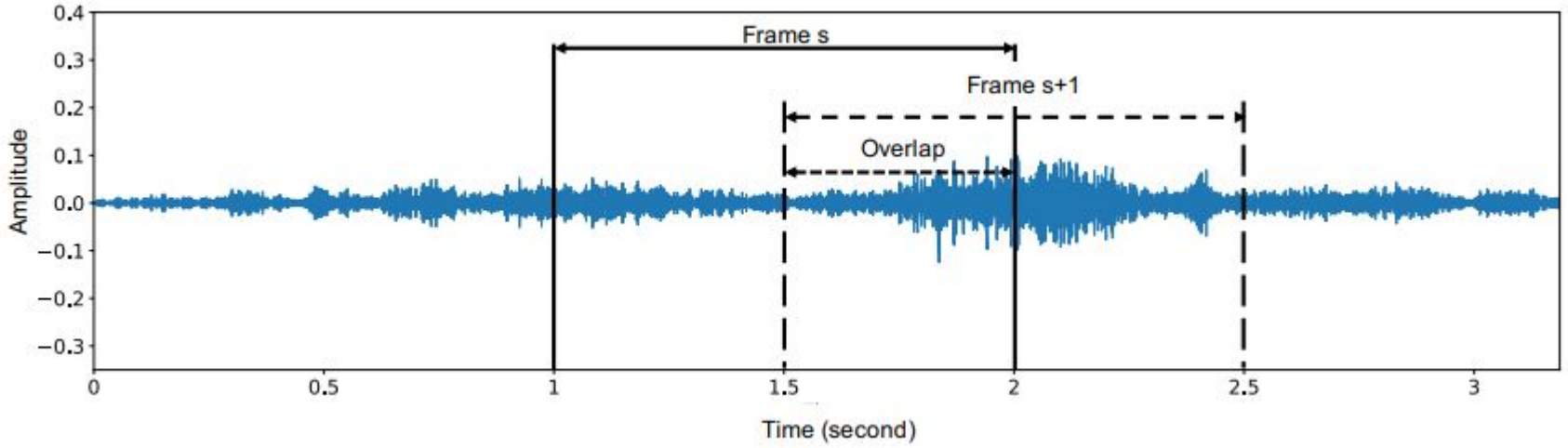
Enfoque de End to End para la clasificación de sonido ambiental basado en una red neuronal de convolucional 1D (CNN) que aprende una representación directamente de la señal de audio.

2. Arquitectura propuesta de End to End

El objetivo de la arquitectura de end to end propuesta es manejar señales de audio de longitudes variables, aprendiendo directamente de la señal de audio, una representación discriminativa que logra un buen rendimiento de clasificación en diferentes sonidos ambientales.

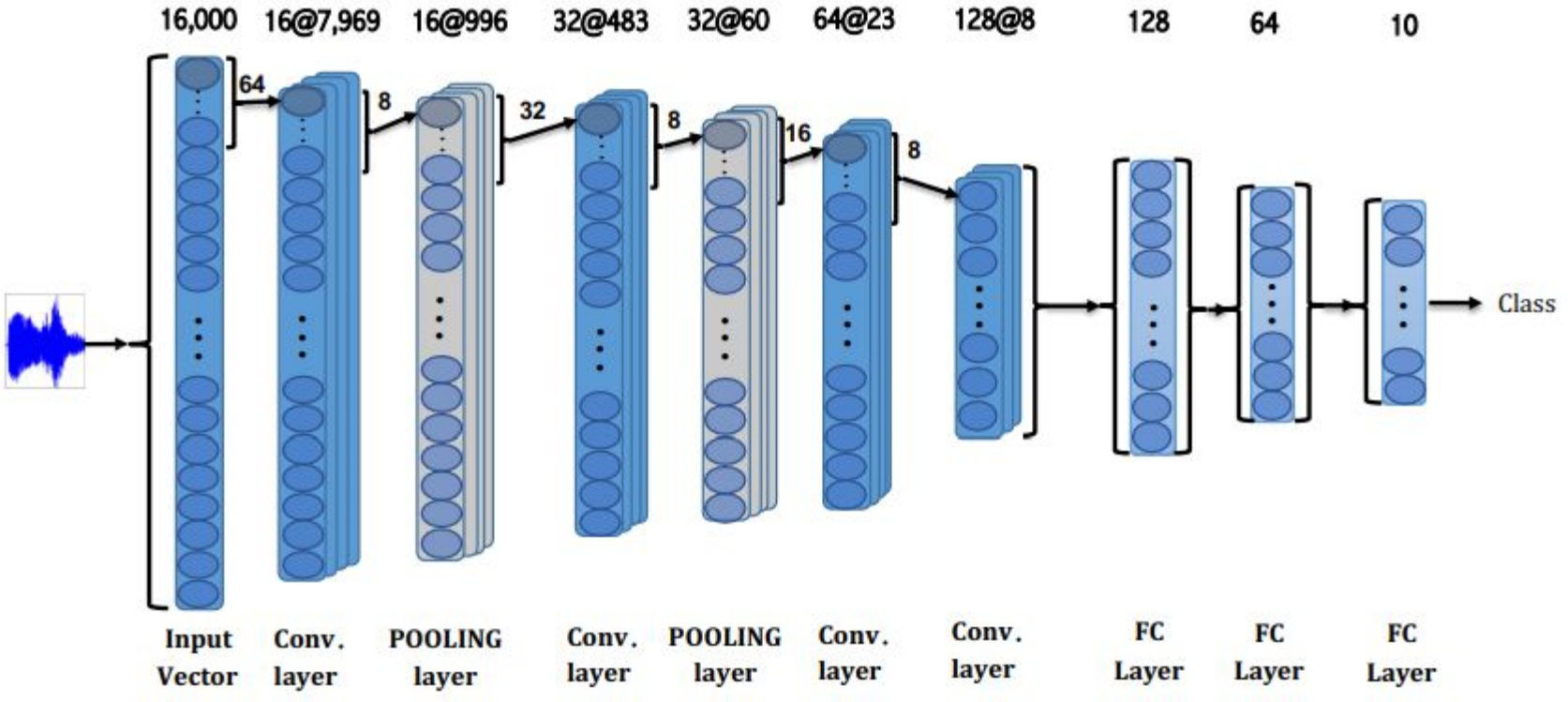
2.1. Longitud de audio variable

Uno de los desafíos de usar CND 1D en el procesamiento de audio es que la longitud de la muestra de entrada debe ser fija.Una forma de evitar esta restricción impuesta por la capa de entrada CNN es dividir la señal de audio en varios cuadros de longitud fija usando una ventana deslizante de ancho apropiado



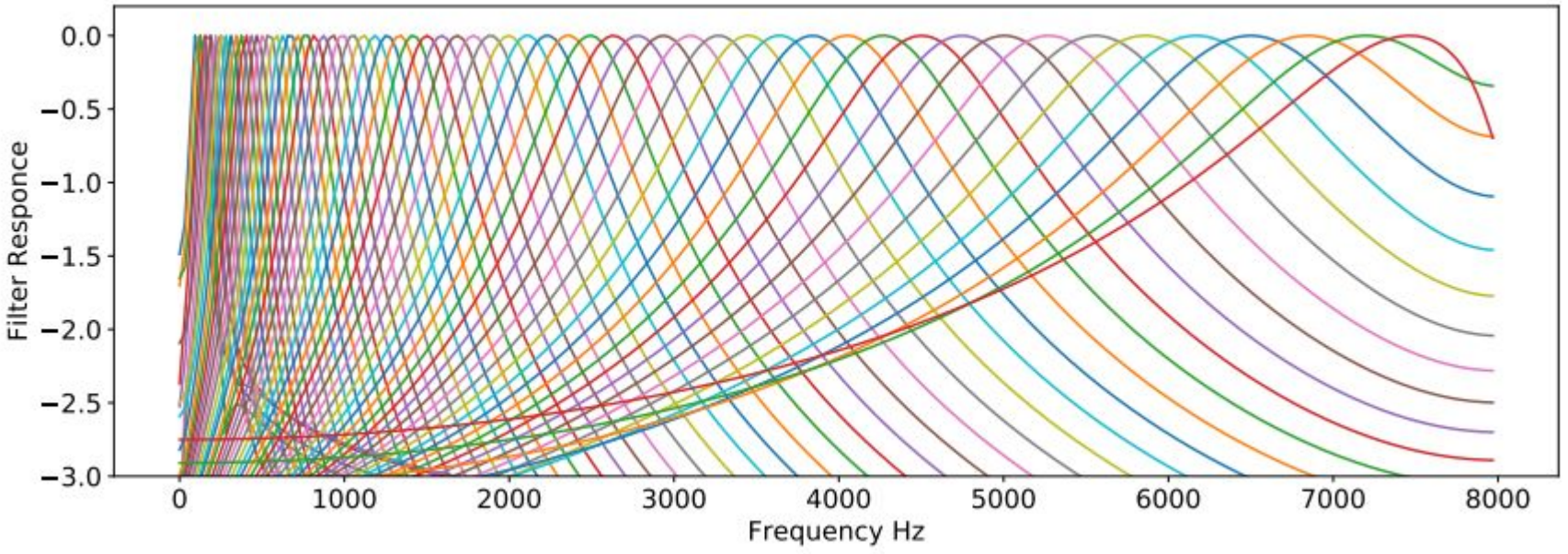
2.2. Topología 1D CNN

La topología propuesta apunta a una arquitectura compacta 1D CNN con un número reducido de parámetros.



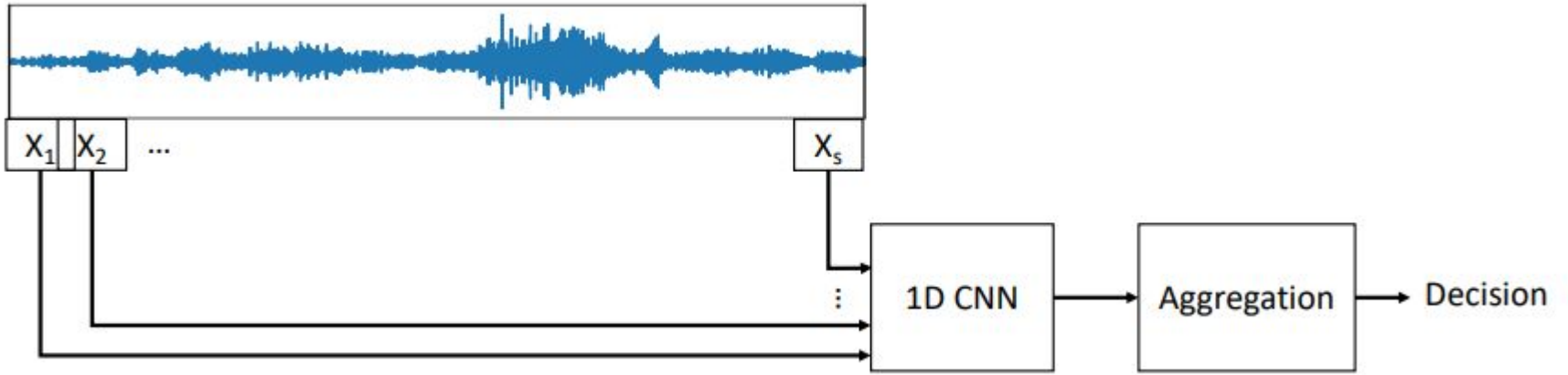
2.3. Gammatone Filterbanks

CND 1D en que su primera capa se puede inicializar como un banco de filtros Gammatone. Un filtro Gammatone es un filtro lineal descrito por una respuesta al impulso de una distribución gamma y un tono sinusoidal.



2.4. Aggregation of Audio Frames

En el caso en que la forma de onda de audio de entrada X se divide en cuadros S indicados como X1, X2,. . . , XS, durante la clasificación necesitamos agregar las predicciones de CNN para llegar a una decisión sobre X



3. Experimental Results

Mejoras en la precisión media para la CND 1D de 16,000 entradas en el conjunto de datos UrbanSound8k.

CL1 Initialization	Window	Overlapping	Combination Rule	Mean Accuracy	# of Parameters
Randomly	Hamming	50%	Sum Rule	83%	256,538
Randomly	Rectangular	50%	Sum Rule	85%	256,538
Randomly	Rectangular	75%	Sum Rule	87%	256,538
Gammatone	Rectangular	50%	Sum Rule	**89%**	550,506

Precisión media de diferentes enfoques en el conjunto de datos UrbanSound8k

Approach	Representation	Mean Accuracy	# of Parameters
Proposed 1D CNN Gamma	1D	**89%**	550 k
Proposed 1D CNN Rand	1D	87%	256 k
SB-CNN (DA) (Salamon & Bello, 2017)	2D	79%	241 k
EnvNet-v2 (Tokozume et al., 2017)	1D	78%	101 M
SKM (DA) (Salamon & Bello, 2015)	2D	76%	NA
SKM (Salamon & Bello, 2015)	2D	74%	NA
PiczakCNN (Piczak, 2015a)	2D	73%	26 M
M18 CNN (Dai et al., 2017)	1D	72%	3.7 M
SB-CNN (Salamon & Bello, 2017)	1D	73%	241 k
VGG (Pons & Serra, 2018)	2D	70%	77 M

NA: Not available. DA: With data augmentation.