Introducción al Aprendizaje Profundo

Berenice & Ricardo Montalvo Lezama

Clasificación de audio

https://github.com/bereml/iap

Abril 2021

Sonido

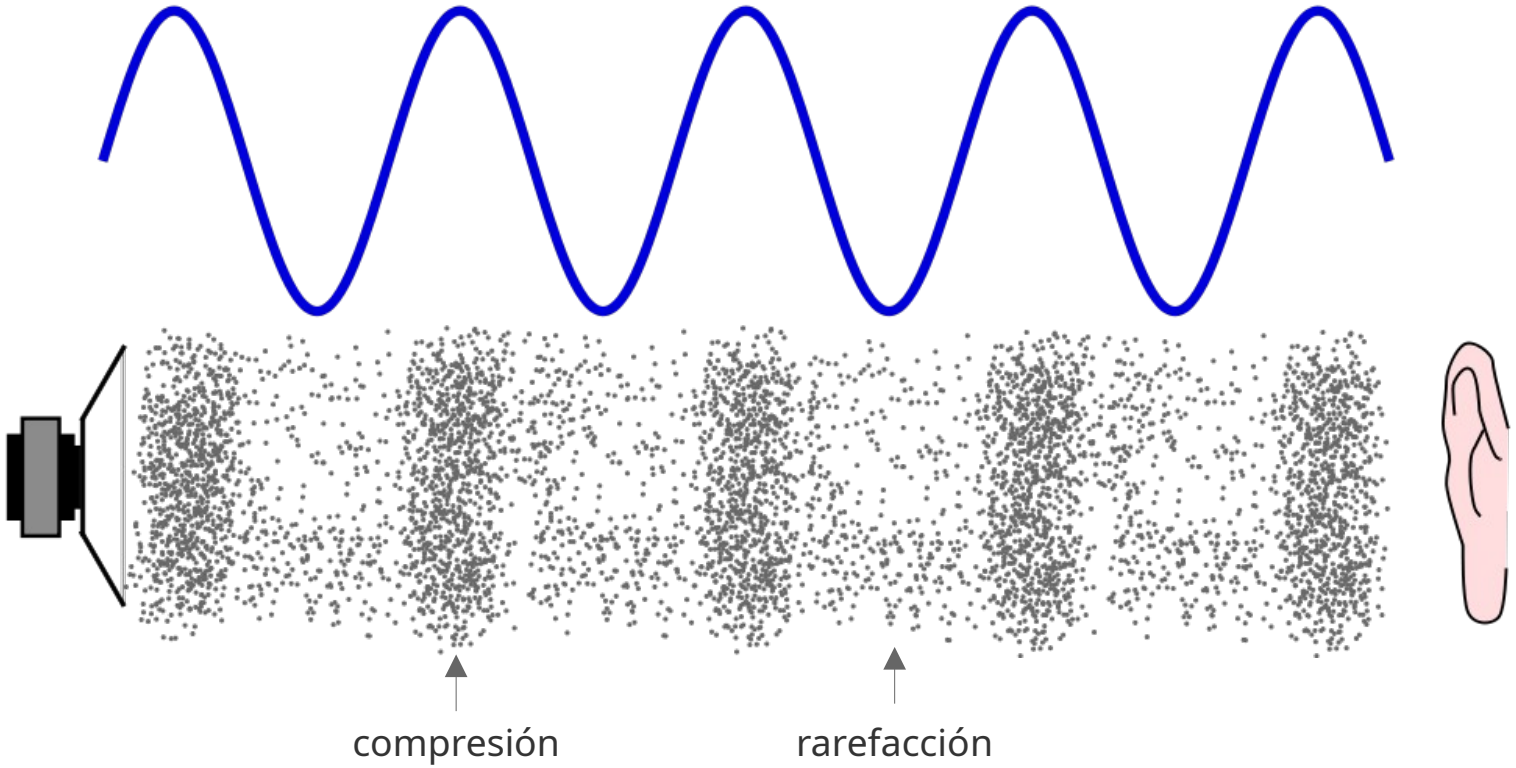
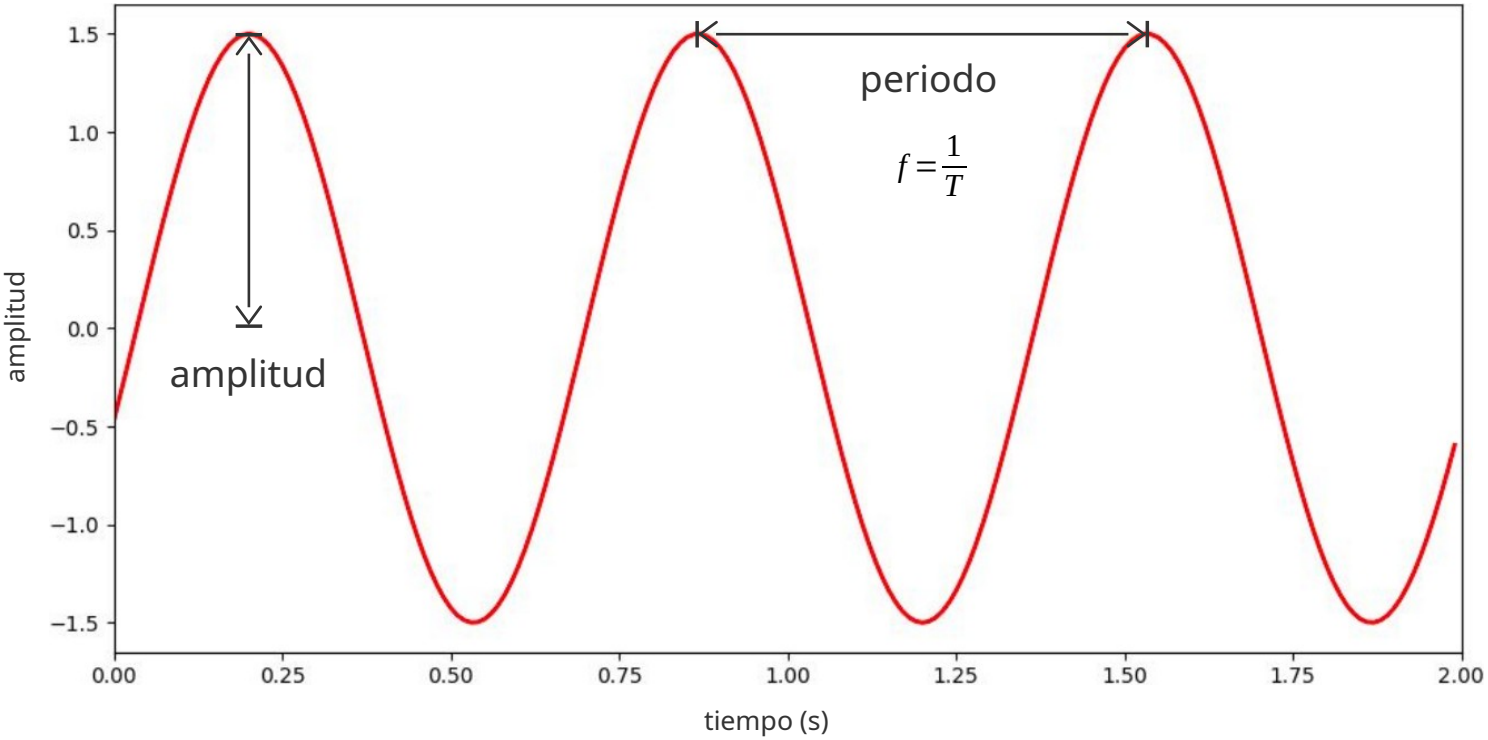
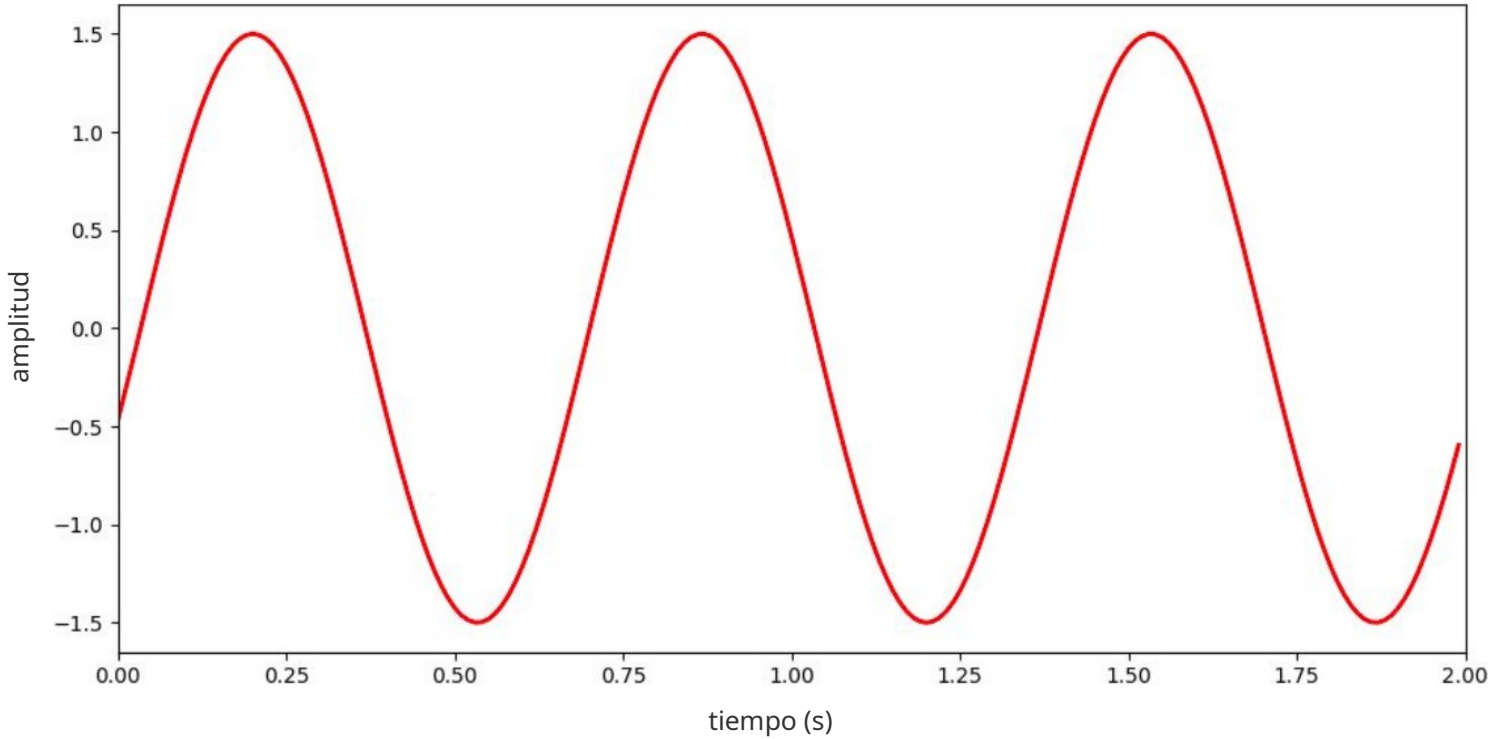


Imagen adaptada de https://kidsdiscover.com/spotlight/sound-and-vibration/

Forma de onda (I)

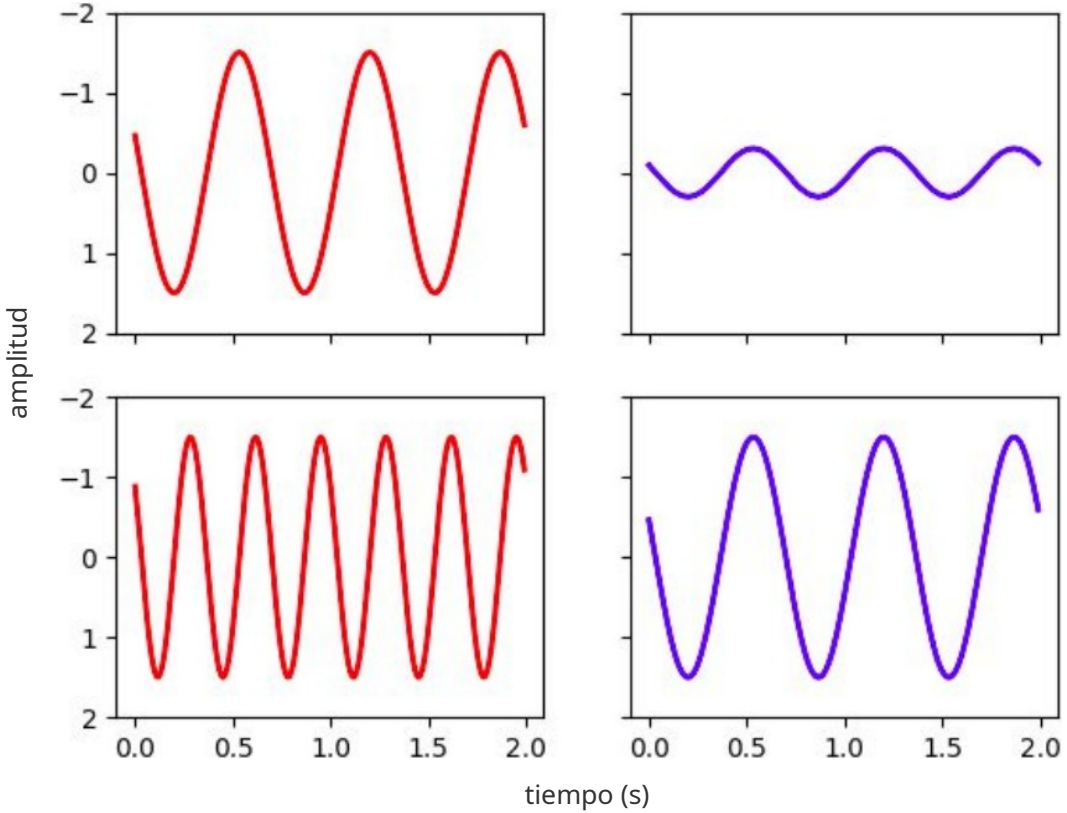


Forma de onda (II)



$$y(t) = A\sin(2\pi f t + \varphi)$$

Amplitud/volumen y frecuencia/tono



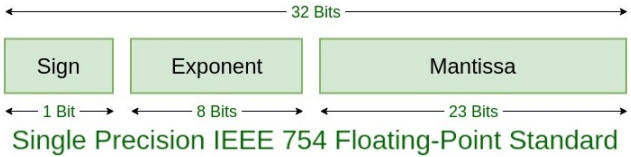
A Physicist's Guide to Music, with some Math thrown in

mayor amplitud → mayor volumen

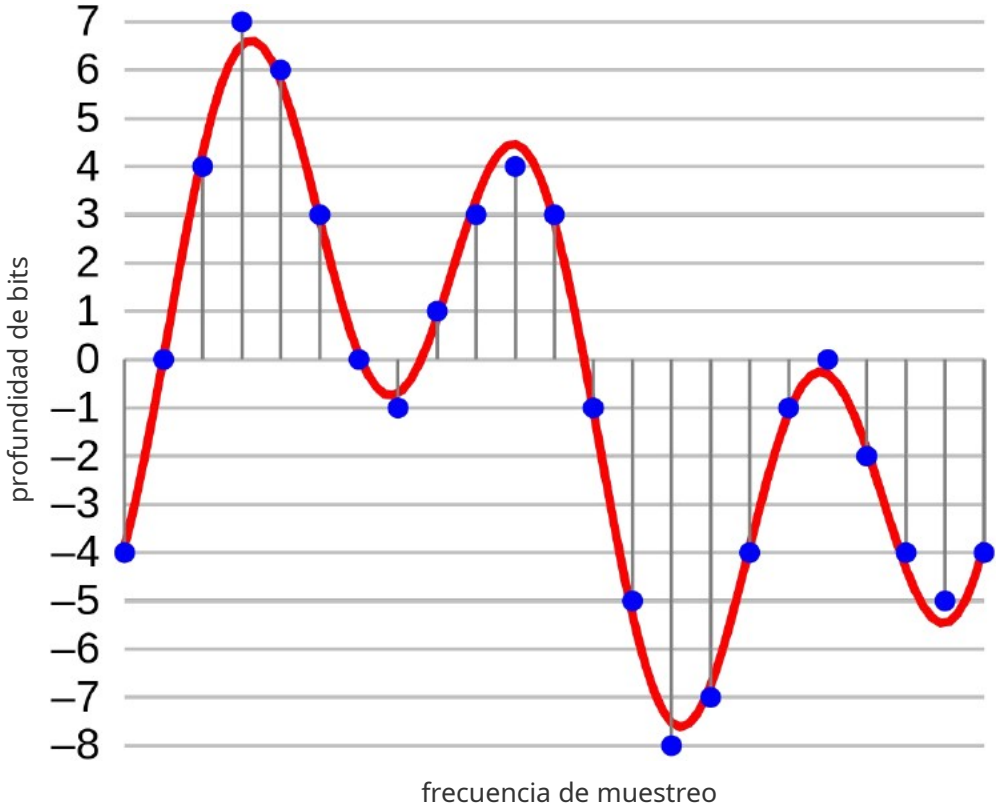
mayor frecuencia → mayor tono

Conversión analógica a digital

- Señales de audio.
  - Continuas y naturales.
- Sistemas digitales.
  - Precisión finita y discreta.



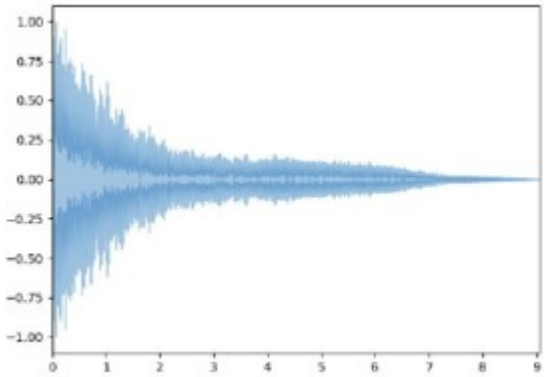
- Digitalización de una señal analógica.
  - Muestreo en frecuencia.
  - Cuantización de las mediciones.



¡tiempo de programar!

`2d_audio_intro.ipynb`

Preprocesamiento de audio para AA



forma de onda

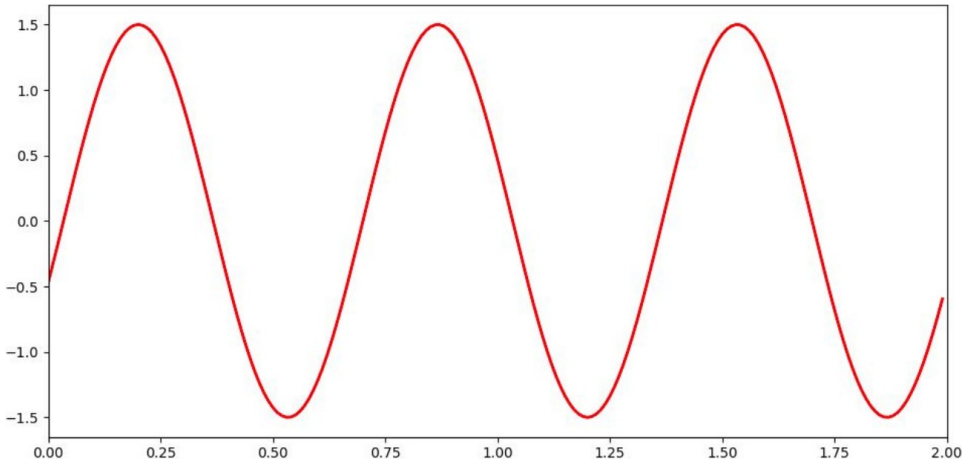
- envolvente de amplitud
- centroide espectral
- tasa de cruce cero
- flujo espectral

ingenieria de características

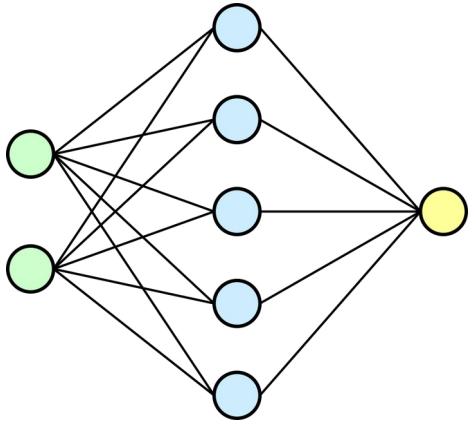
- regresión lineal
- SVM
- HMM
- k-medias

algoritmo AA

AP empleando audio en crudo (I)

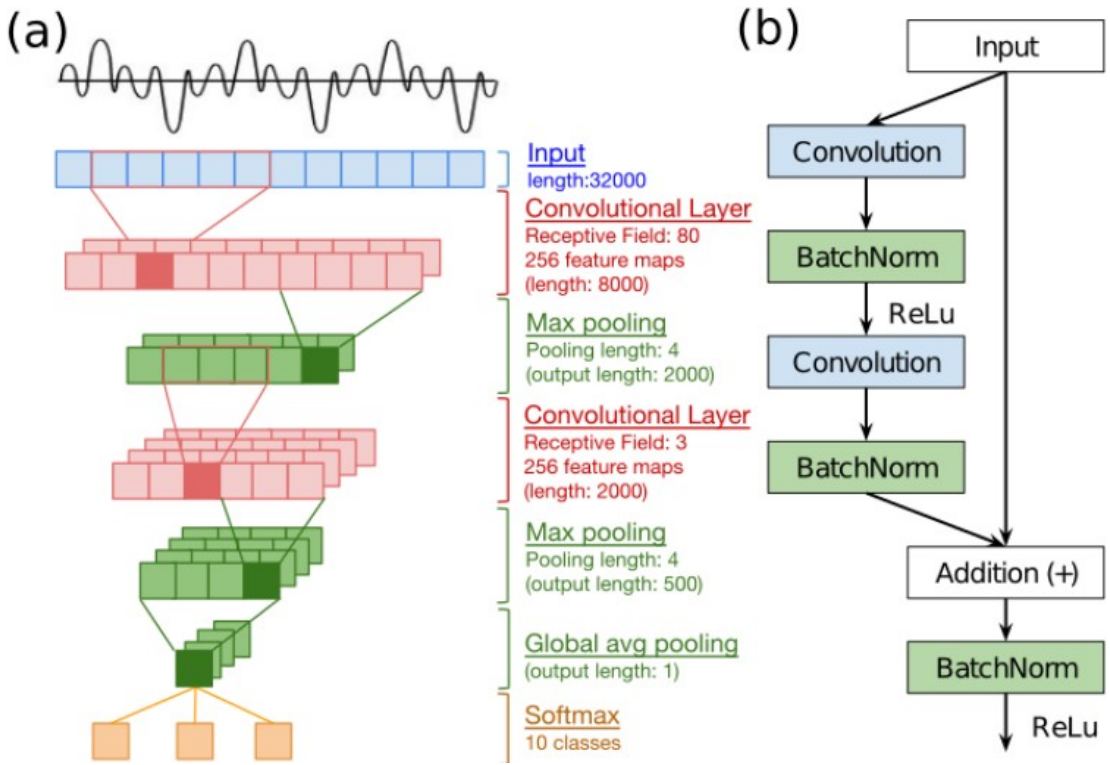


forma de onda



red neuronal

AP empleando audio en crudo (II)



M3 (0.2M)	M5 (0.5M)	M11 (1.8M)	M18 (3.7M)	M34-res (4M)
Input: 32000x1 time-domain waveform				
[80/4, 256]	[80/4, 128]	[80/4, 64]	[80/4, 64]	[80/4, 48]
Maxpool: 4x1 (output: 2000 × n)				
[3, 256]	[3, 128]	[3, 64] × 2	[3, 64] × 4	$\begin{bmatrix} 3, 48 \\ 3, 48 \end{bmatrix} \times 3$
Maxpool: 4x1 (output: 500×n)				
	[3, 256]	[3, 128] × 2	[3, 128] × 4	$\begin{bmatrix} 3, 96 \\ 3, 96 \end{bmatrix} \times 4$
	Maxpool: 4x1 (output: 125 × n)			
	[3, 512]	[3, 256] × 3	[3, 256] × 4	$\begin{bmatrix} 3, 192 \\ 3, 192 \end{bmatrix} \times 6$
	Maxpool: 4x1 (output: 32 × n)			
		[3, 512] × 2	[3, 512] × 4	$\begin{bmatrix} 3, 384 \\ 3, 384 \end{bmatrix} \times 3$
Global average pooling (output: 1 × n)				
Softmax				

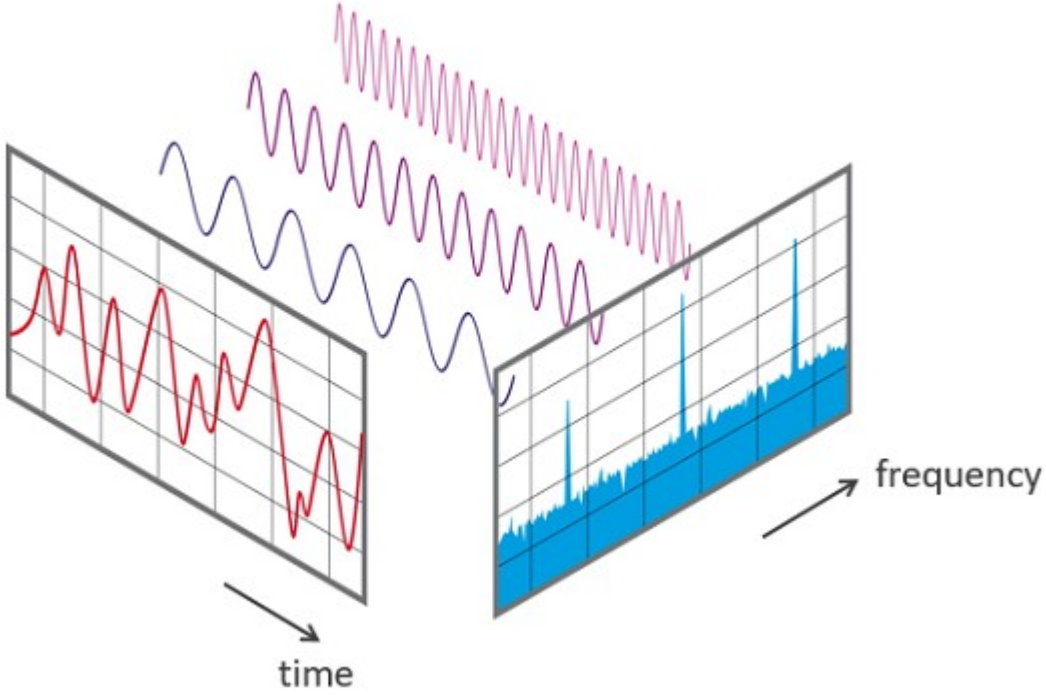
Dai et. al. Very Deep Convolutional Neural Networks For Raw Waveforms. 2017.
https://github.com/engmubarak48/Very_Deep_CNN_AudioClassification

¡tiempo de programar!

`2e_speech_cnn1d.ipynb`

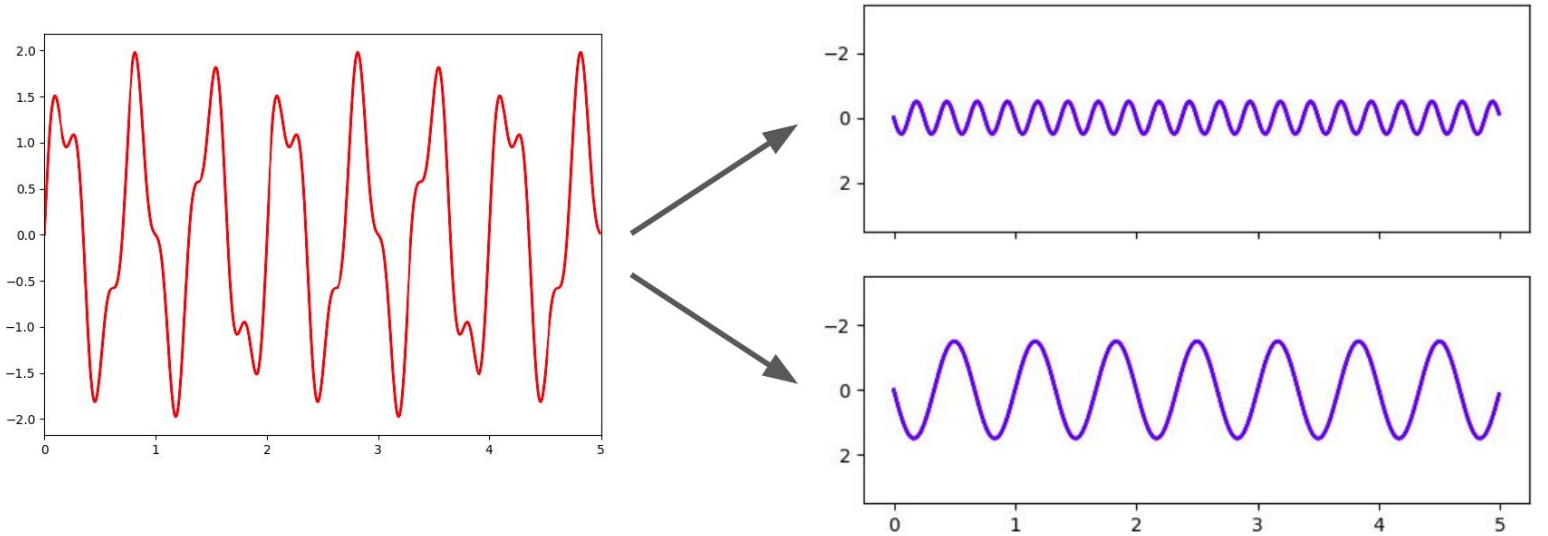
Transformada de Fourier (I)

- Descomponer una señal periódica compleja en la suma de funciones seno a diferentes frecuencias.



Transformada de Fourier (II)

- Descomponer una señal periódica compleja en la suma de funciones seno a diferentes frecuencias.



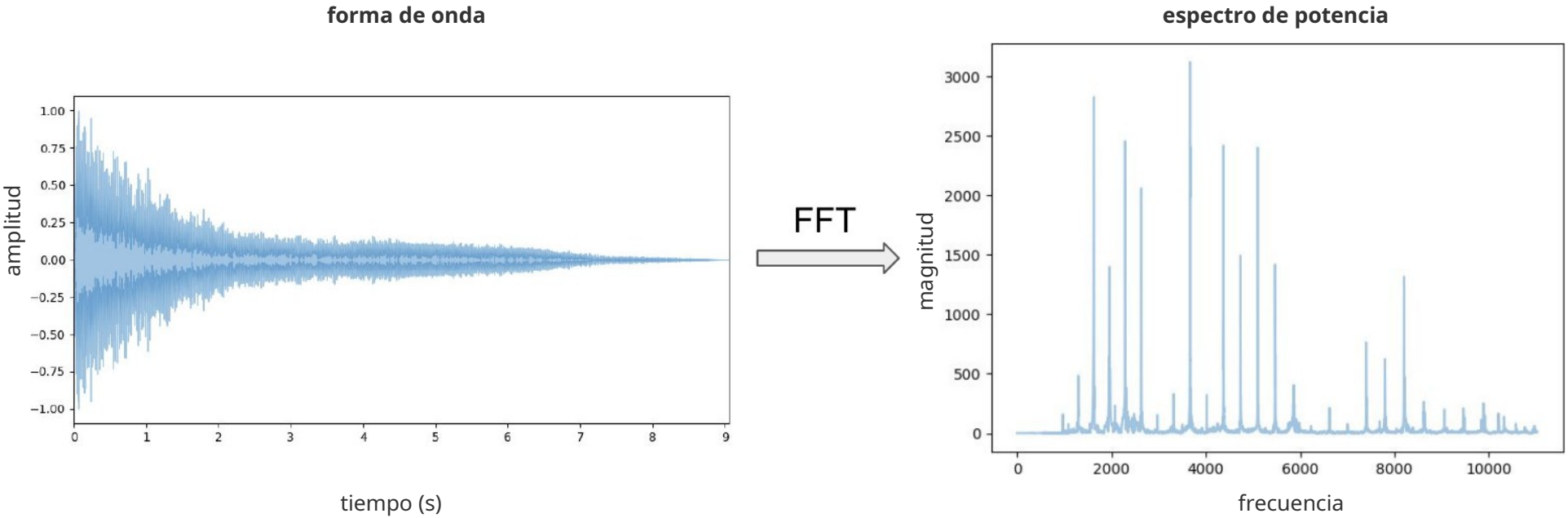
$A_1 = 0.5 \quad f_1 = 4.0 \quad \varphi_1 = 0$

$A_1 = 1.5 \quad f_1 = 1.5 \quad \varphi_1 = 0$

$$y(t) = A_1 \sin(2\pi f_1 t + \varphi_1) + A_2 \sin(2\pi f_2 t + \varphi_2)$$

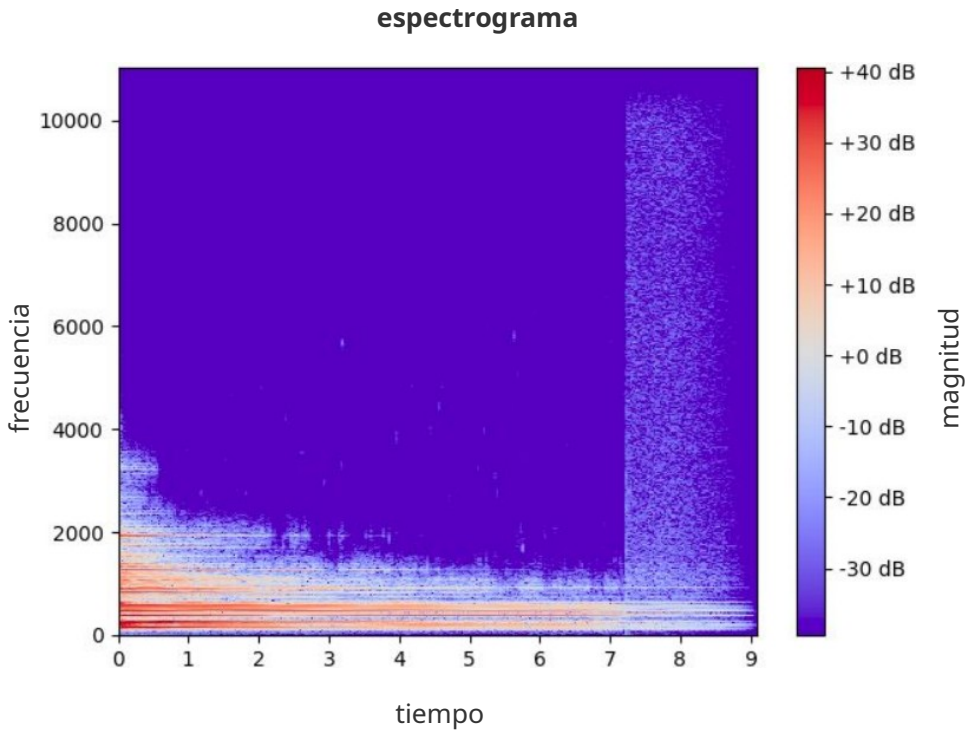
Espectro de potencia

- Transformación del dominio del tiempo al dominio frecuencia.
- Elimina información temporal.

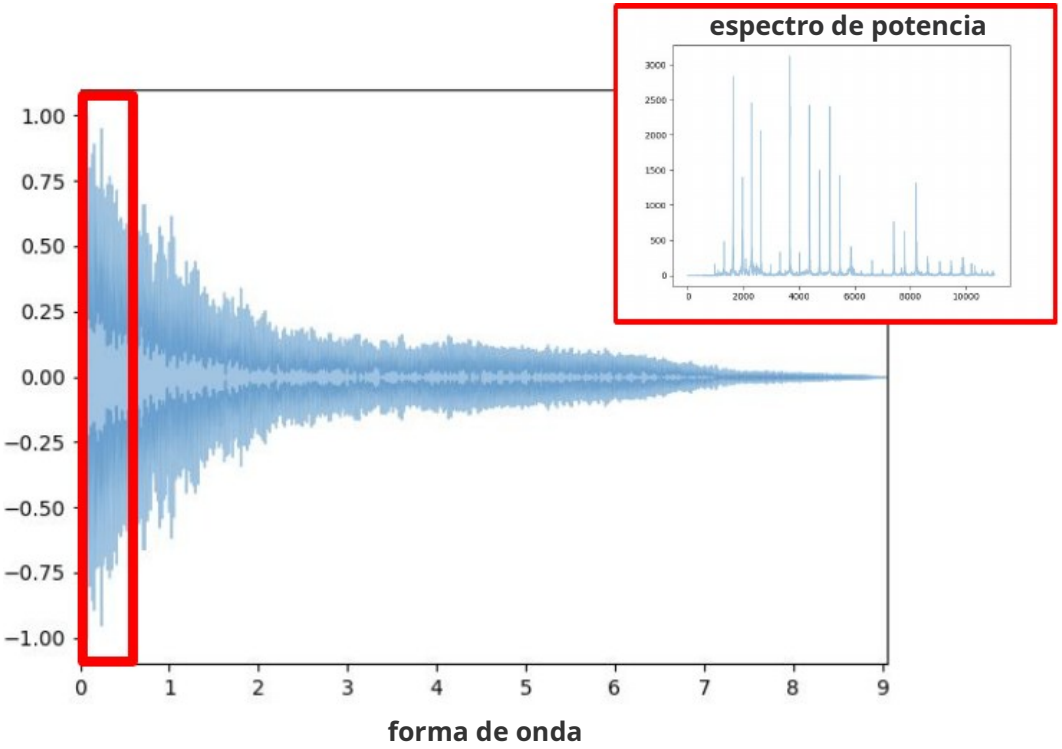


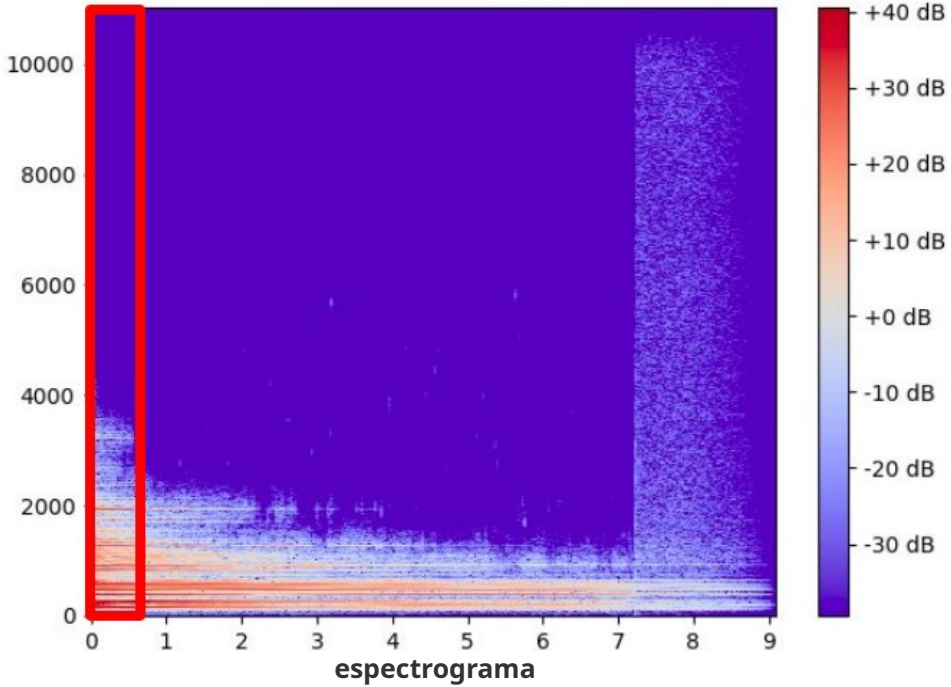
Espectrograma

- Computamos varias FFT en diferentes intervalos.
- Preserva información temporal.
- Intervalos (ventanas) de tamaño fijo.
- Espectrograma: tiempo + frecuencia + magnitud.

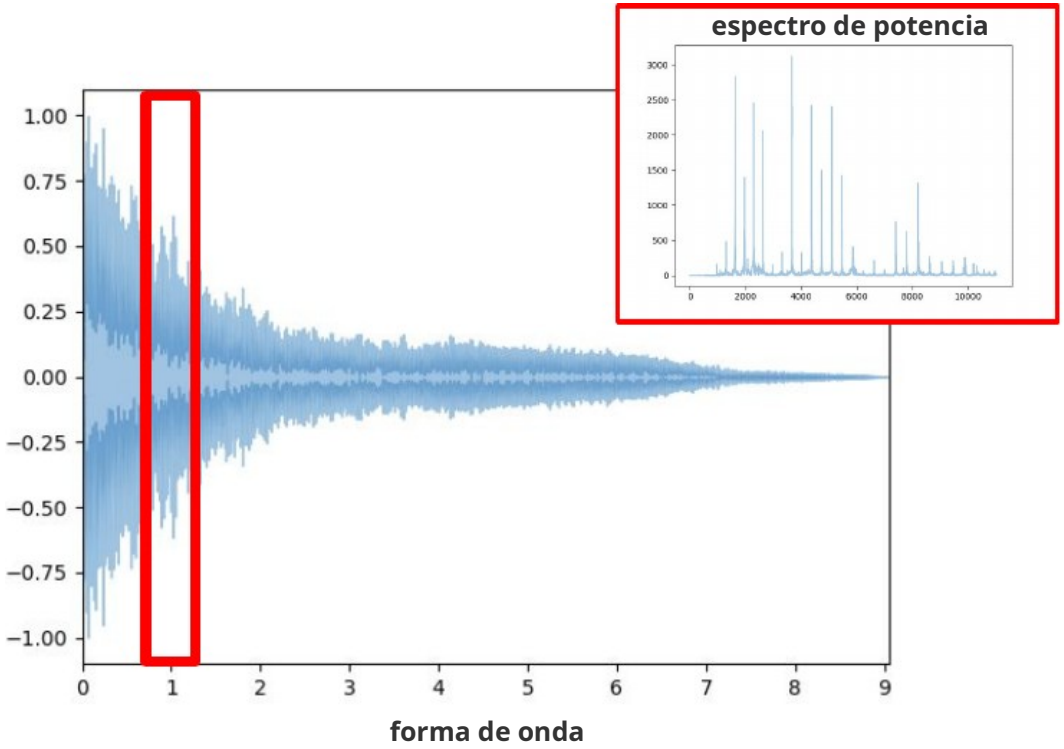


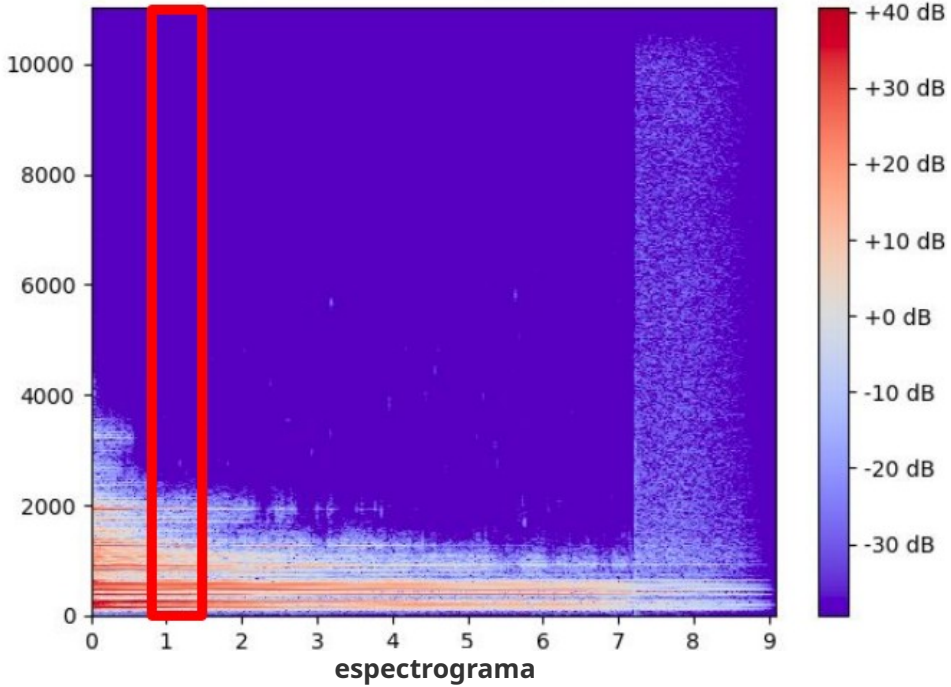
Transformada de Fourier de Tiempo Corto (I)



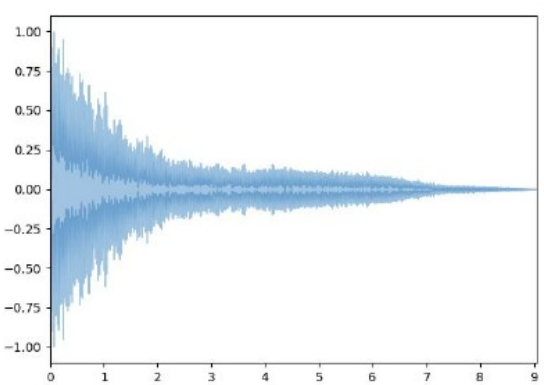


Transformada de Fourier de Tiempo Corto (II)

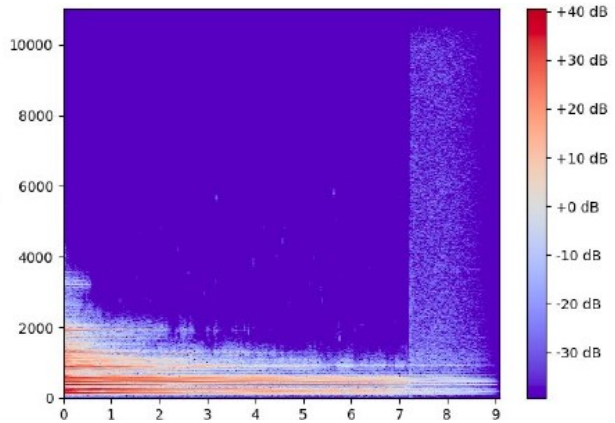




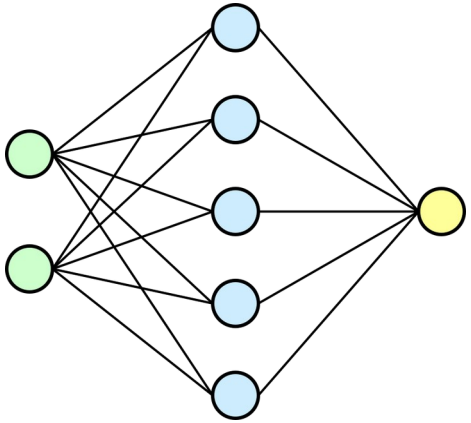
Espectrogramas para AP



forma de onda



espectrograma



red neuronal

Coeficientes Cepstrales en Frecuencias de Mel - MFCC

- Espectograma => MFCC.

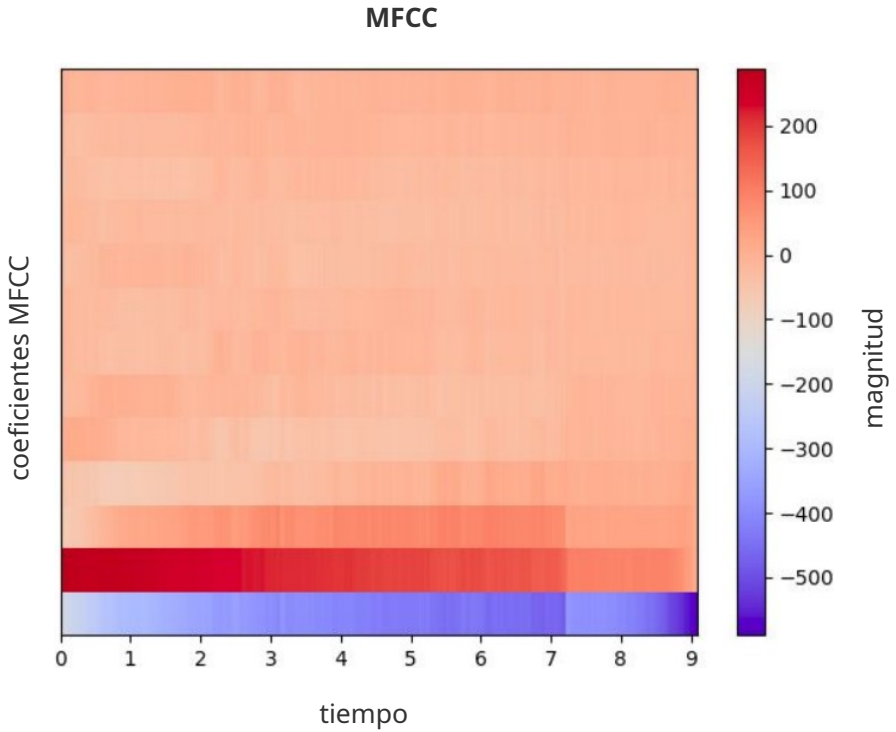
  1) Aplicamos bancos de filtros en escala de Mel y obtenemos espectograma de Mel.

  2) Aplicamos tranformada DCT para obtener MFCCs.

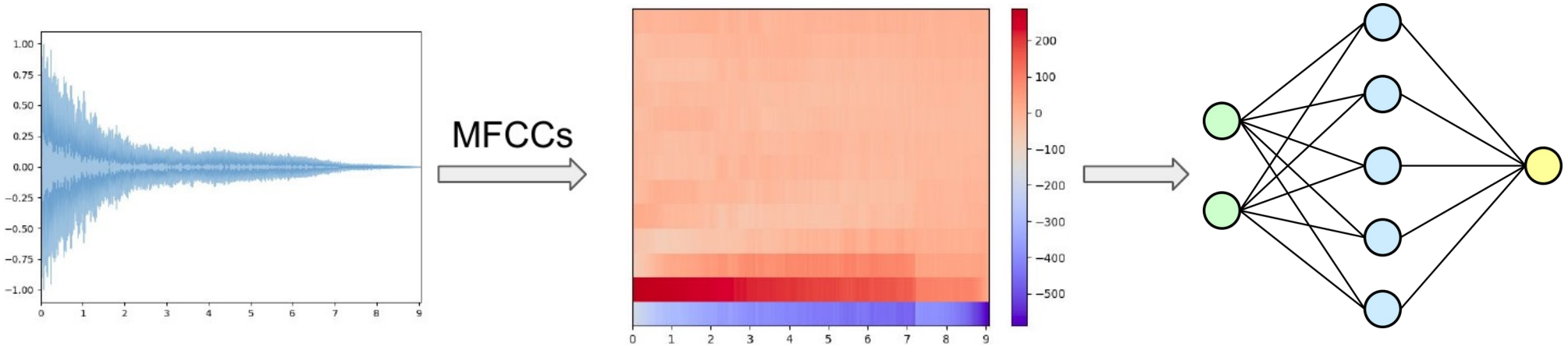
- Preserva información temporal.

- Aproximan la percepción humana de las frecuencias (logarítmica).

- Espectrograma: tiempo + frecuencia + magnitud.



MFCCs para AP



forma de onda **MFCCs** **red neuronal**

1. M. Huzaifah. Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks. 2017.
2. Slovyev. Deep Learning Approaches for Understanding Simple Speech Commands. 2018.

¡tiempo de programar!

`2f_speech_cnn2d.ipynb`