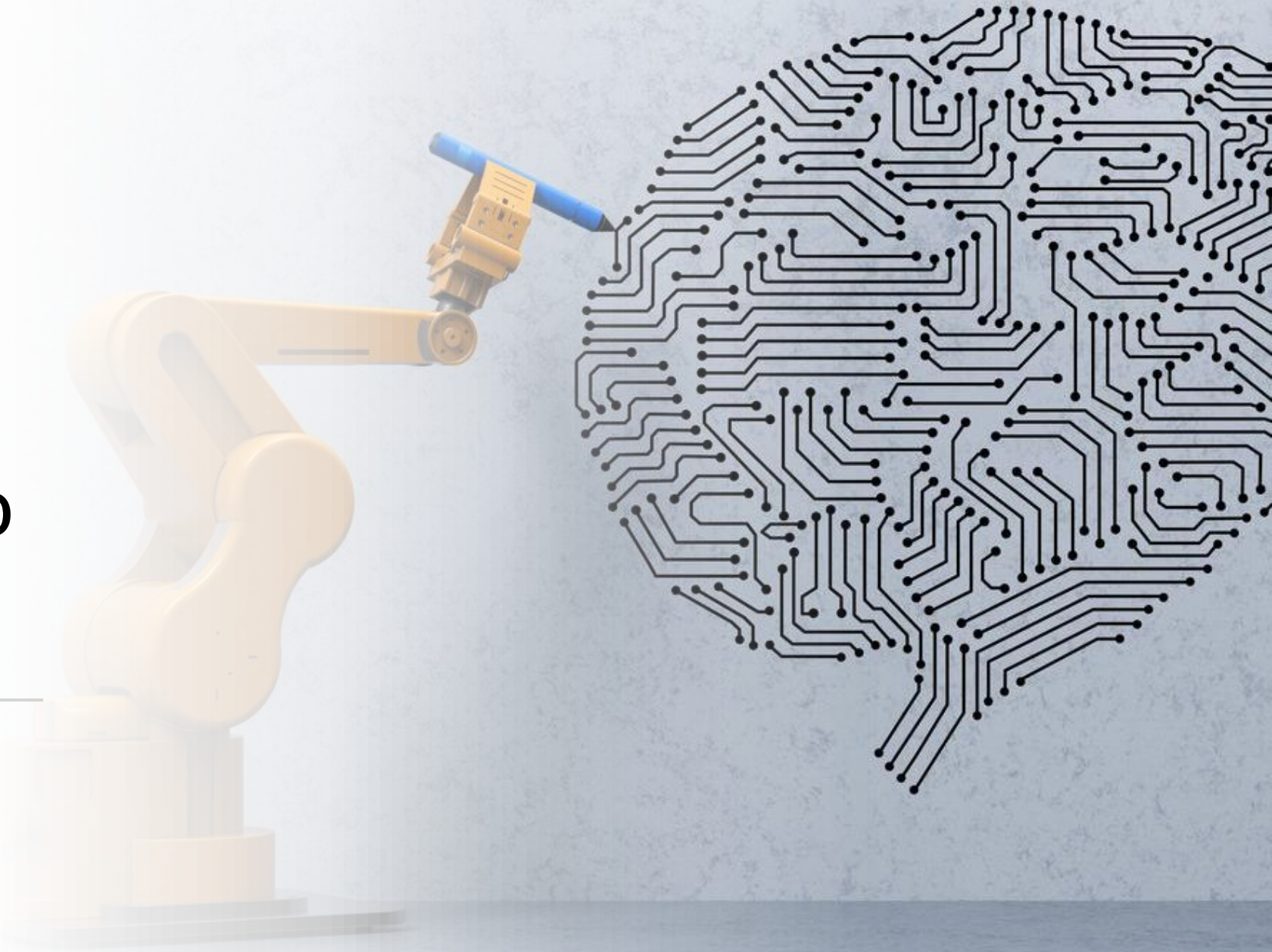




Reconocimiento de Patrones

Clase 4

Extracción de características III



Antes de empezar

- ¿Dudas de la clase anterior?



Histogram of Oriented Gradients (HOG)

- Es un descriptor de características ampliamente utilizado, el cual representa la forma de un objeto codificando los gradientes de intensidad y sus orientaciones dentro de regiones localizadas de una imagen.
- HOG captura estructuras locales de bordes y gradientes, lo que permite a las máquinas interpretar datos visuales de manera efectiva.
- Excelente para la detección humana y en otros problemas de detección de objetos.

Input image

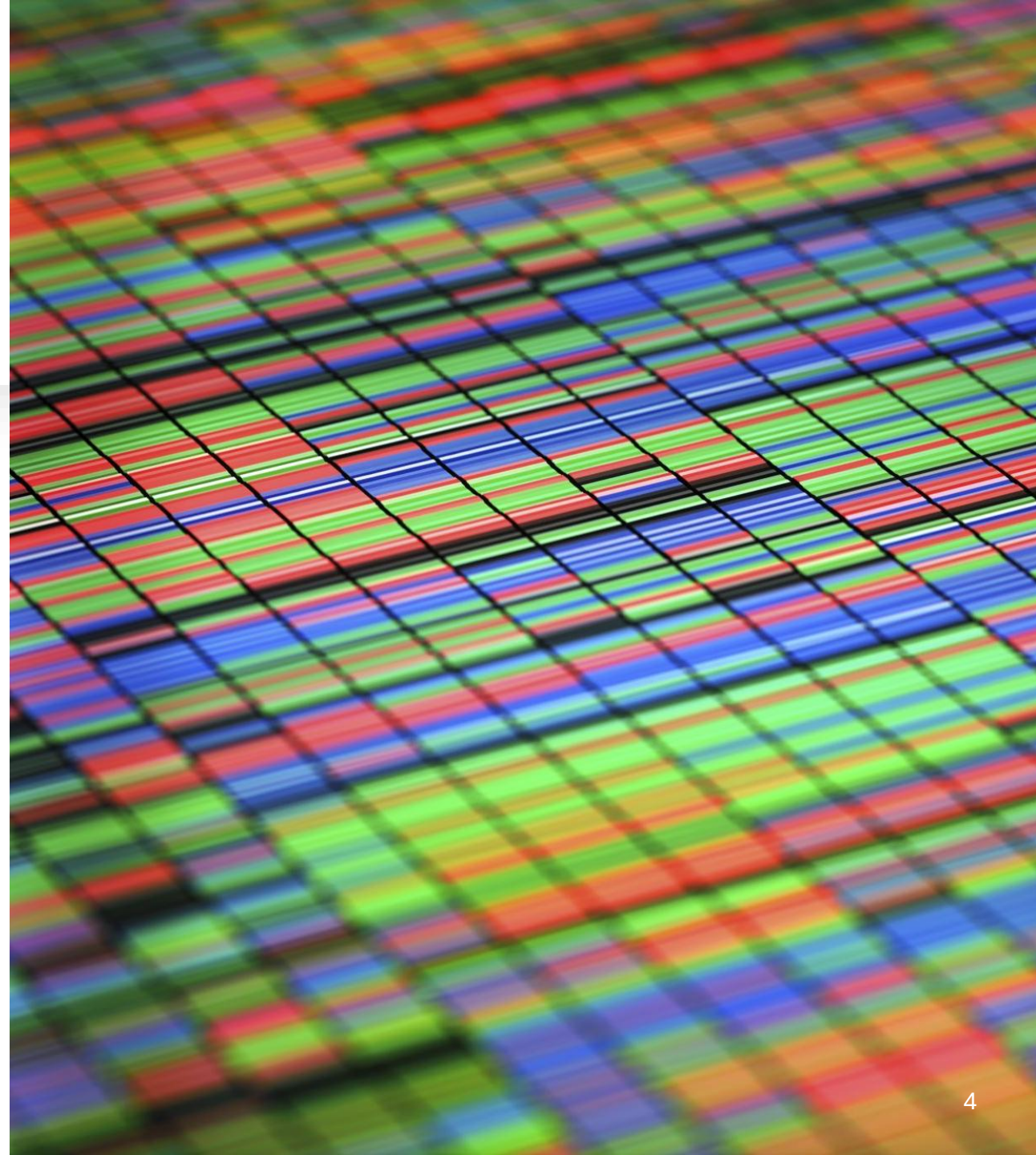


Histogram of Oriented Gradients



Descripción general del algoritmo

1. Preprocesamiento.
2. Calculo de los gradientes de la imagen: horizontal (G_x) y vertical (G_y).
3. Calculo de los histogramas de gradientes.
4. Normalización entre bloques.
5. Concatenación de histogramas en un vector de características.



1. Preprocesamiento.



Imagen Original
1090 x 612

Recortar



Imagen Recortada
196 x 438

Redimensionar



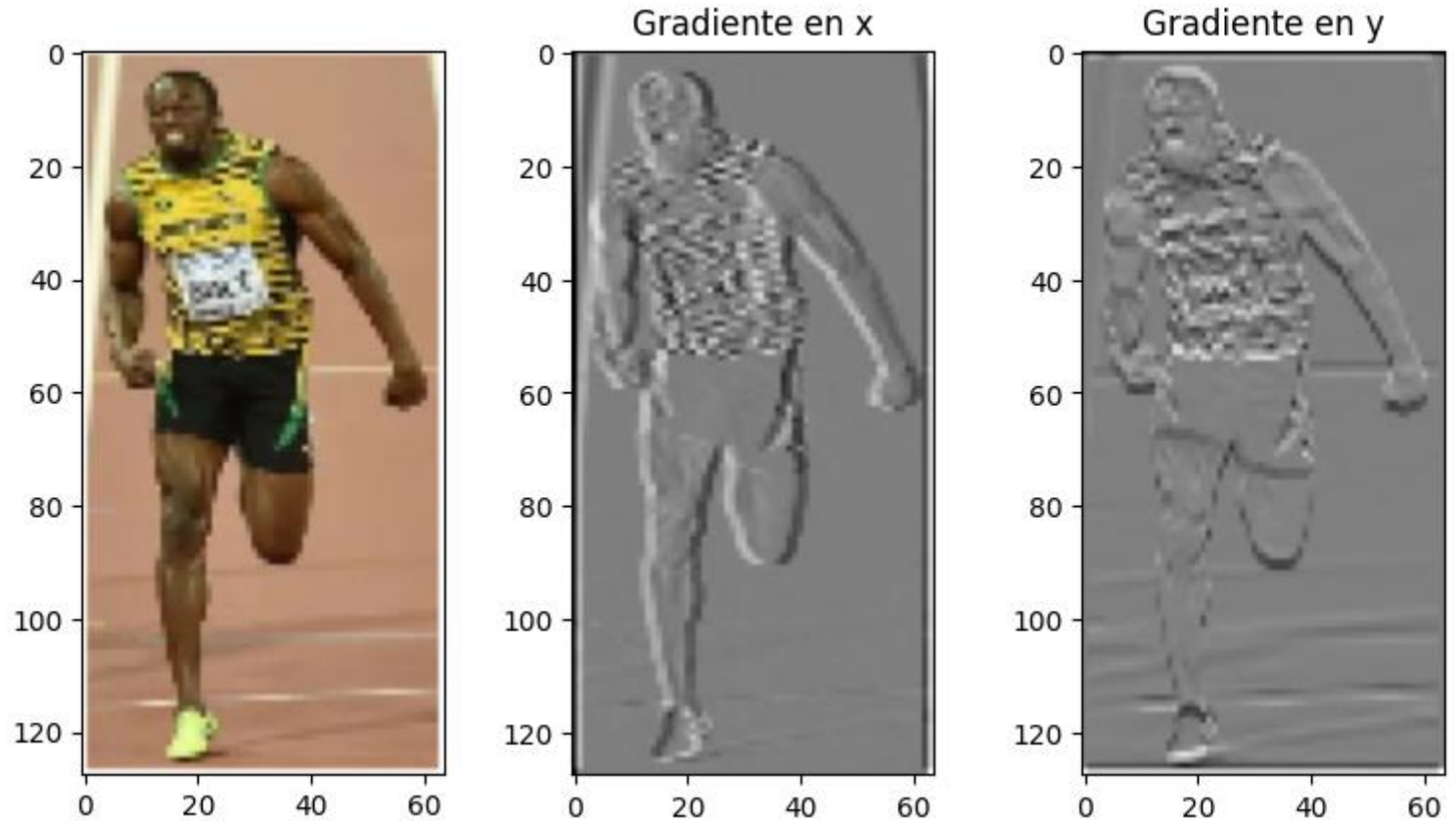
Imagen
Redimensionada
64 x 128

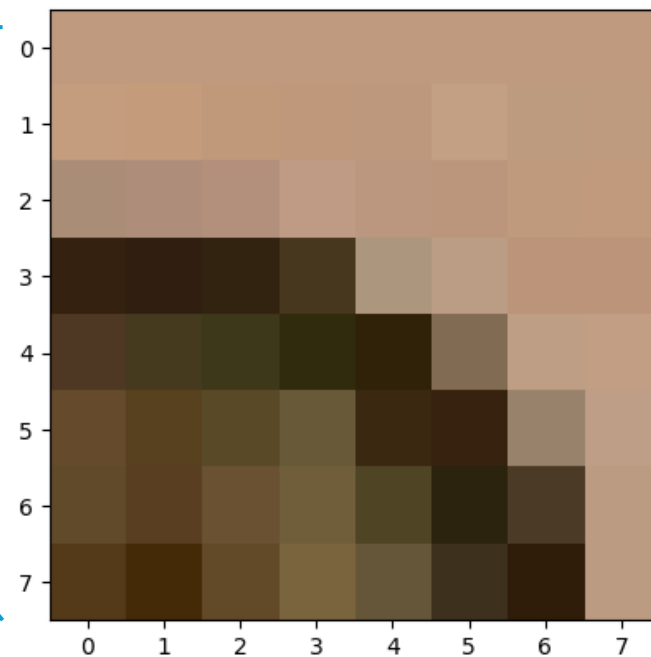
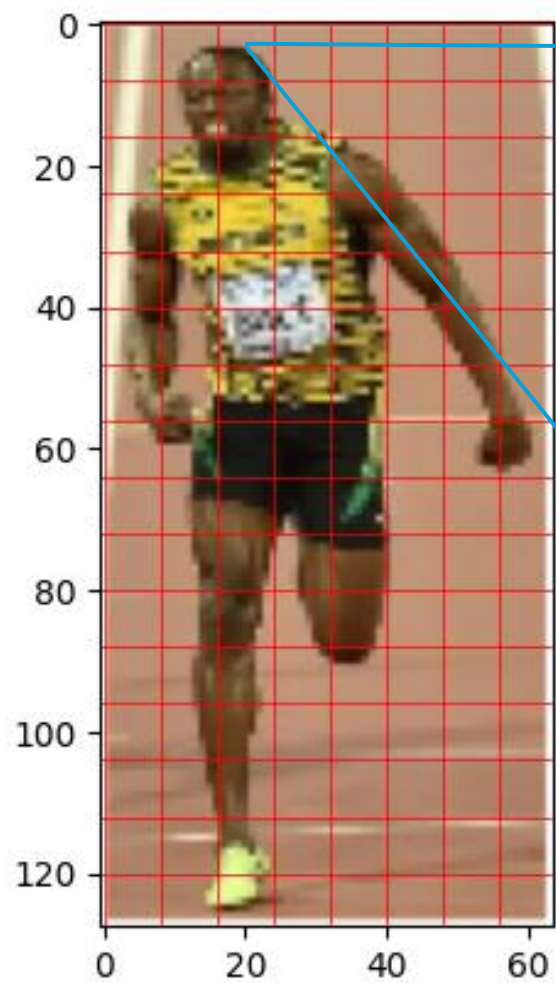
2. Calculo de los gradientes

$$G_x = I * [-1, 0, 1] \quad G_y = I * [-1, 0, 1]^T$$

$$\text{Magnitud: } M = \sqrt{G_x^2 + G_y^2}$$

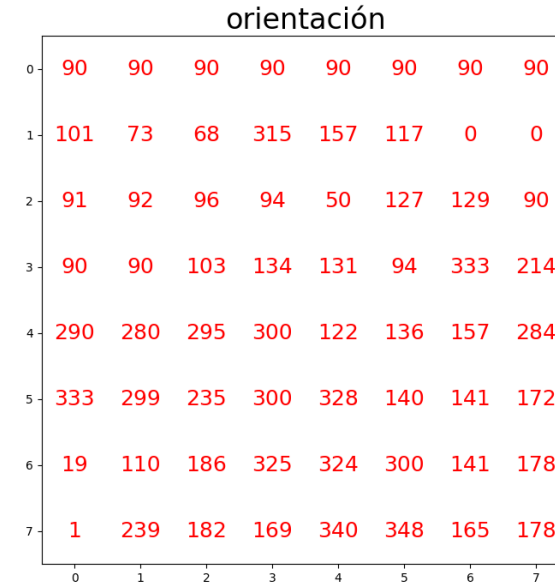
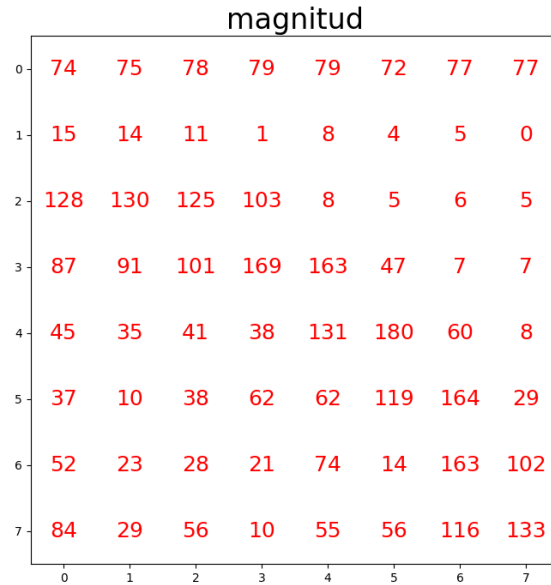
$$\text{Orientación: } \theta = \tan^{-1} \left(\frac{G_y}{G_x} \right)$$





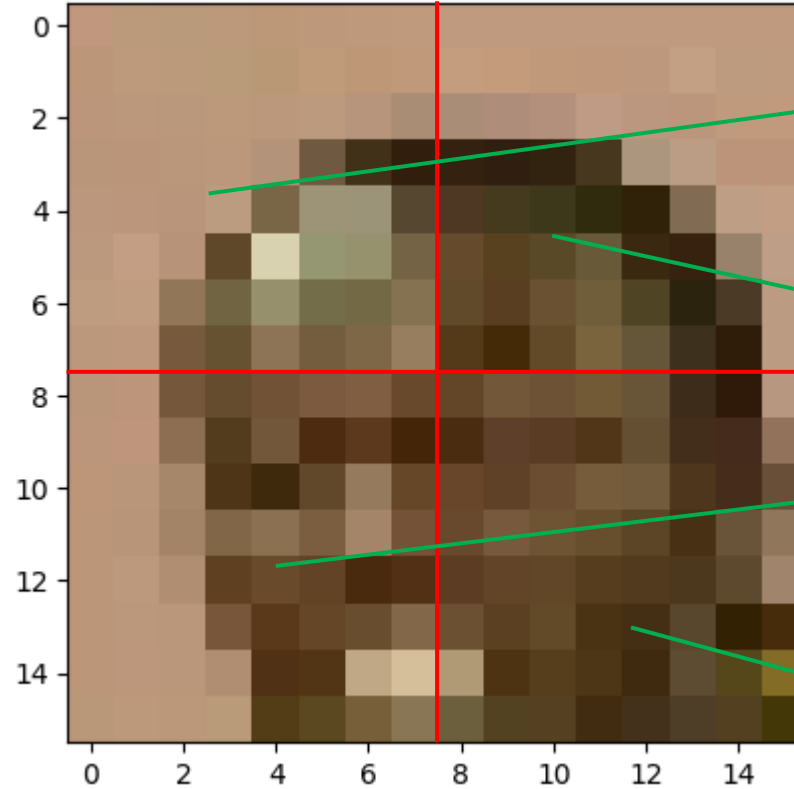
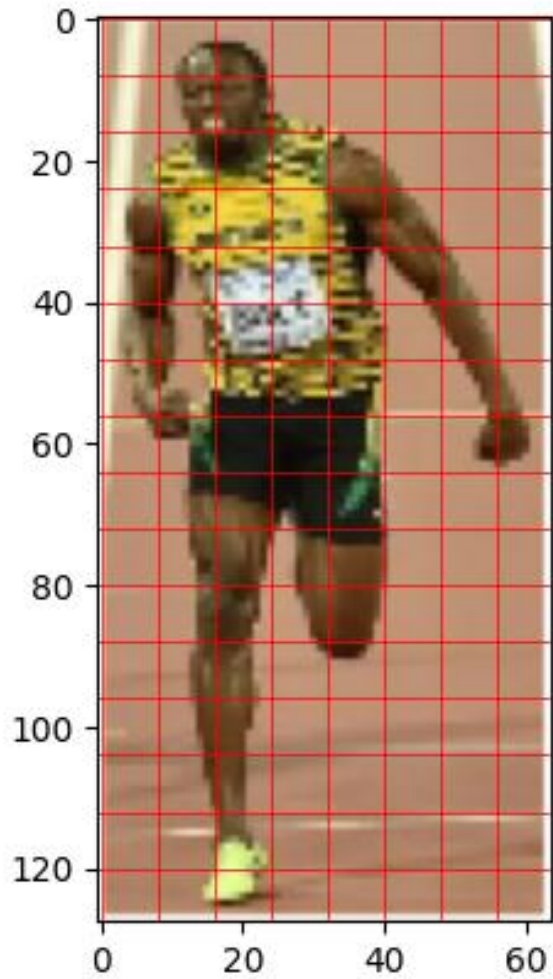
0	162	162	162	162	162	162	162
1	165	164	161	160	160	167	162
2	147	149	152	163	159	158	162
3	37	34	37	57	154	163	157
4	60	58	54	41	35	111	165
5	79	68	74	90	43	38	134
6	77	67	85	95	68	35	61
7	62	46	77	102	87	50	32

3. Calculo de histogramas de gradientes



0	20	40	60	80	100	120	140	160
---	----	----	----	----	-----	-----	-----	-----

4. Normalización entre bloques



50	80	15	60	71	120	101	42	4
----	----	----	----	----	-----	-----	----	---

15	61	65	90	150	110	50	80	60
----	----	----	----	-----	-----	----	----	----

6	54	78	112	52	140	163	180	10
---	----	----	-----	----	-----	-----	-----	----

170	190	200	90	172	56	167	98	67
-----	-----	-----	----	-----	----	-----	----	----

$$v = [v_1, v_2, \dots, v_{36}]$$

- Norma L2

$$f = \frac{v}{\sqrt{\|v\|_2^2 + e}}$$

- Norma L1

$$f = \frac{v}{(\|v\|_1 + e)}$$

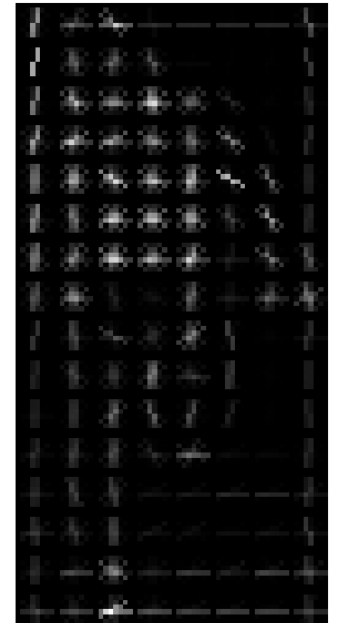
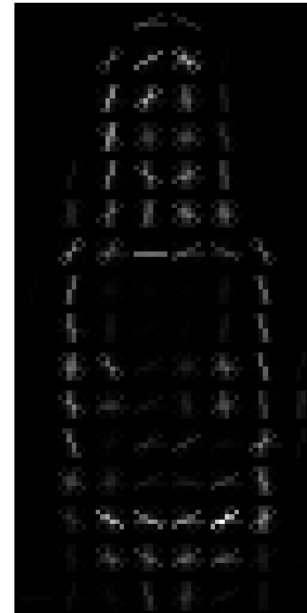
- L1-sqr

$$f = \sqrt{\frac{v}{(\|v\|_1 + e)}}$$

5. Concatenación de histogramas.

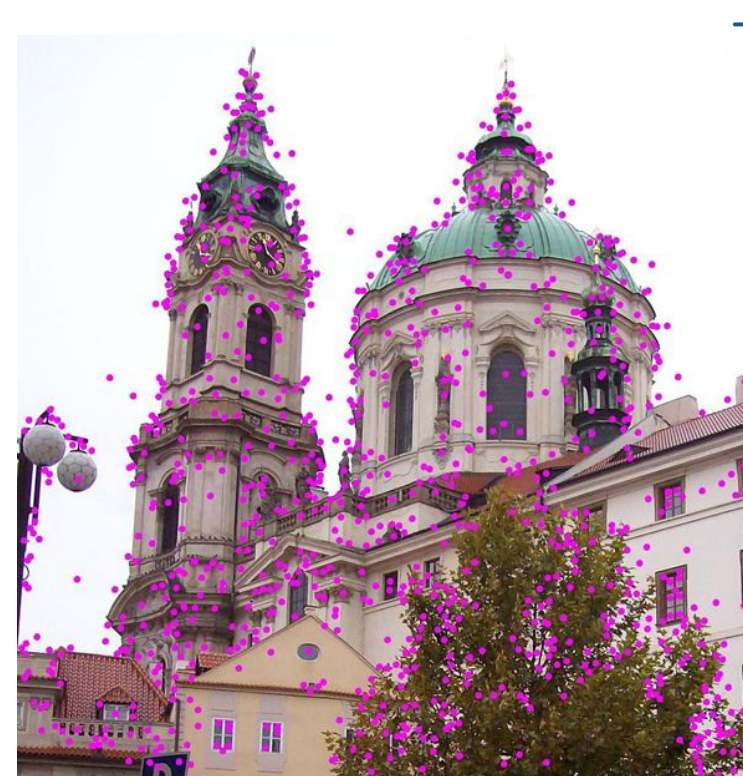
Para este caso, se tiene un vector de características de 3780 valores.

$$f = [f_1, f_2, \dots, f_{3780}]$$



Scale-invariant feature transform (SIFT)

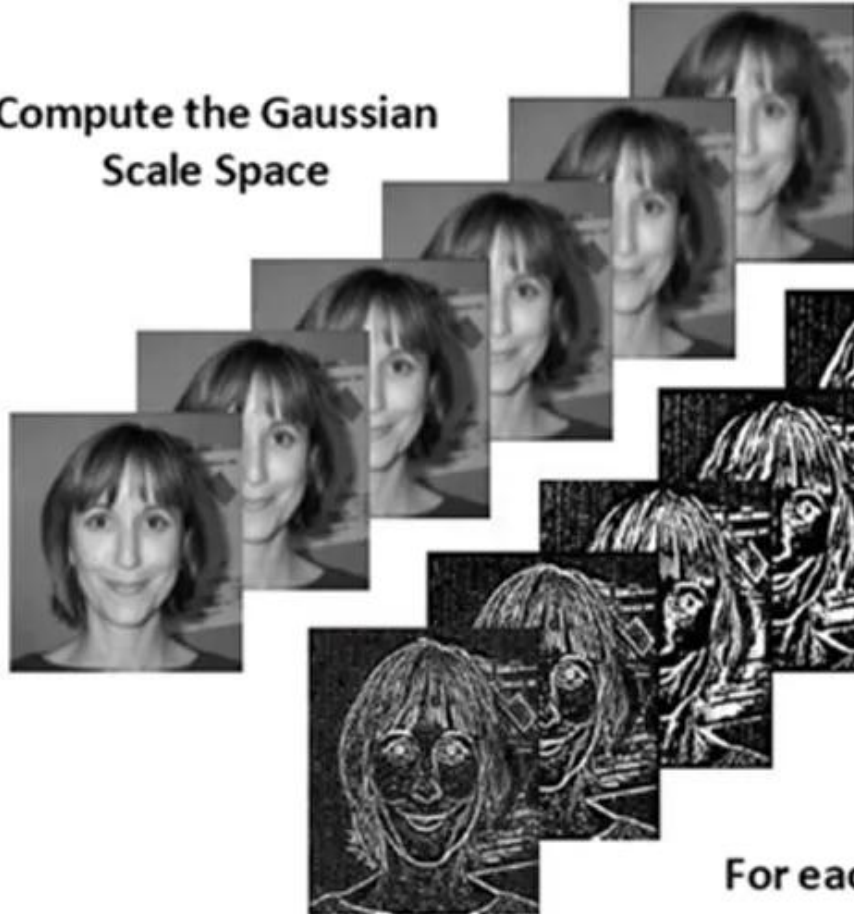
- Es un algoritmo que detecta y describe características locales en imágenes que son invariantes a:
 - Escala
 - Rotación
 - Cambios en la iluminación
 - Cambios moderados de perspectiva.
- Este algoritmo puede identificar de forma fiable objetos en imágenes, sin importar como se hayan capturado.



Source image (640x480) pixel

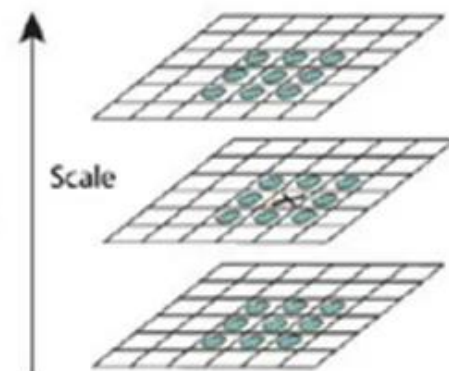


Compute the Gaussian Scale Space

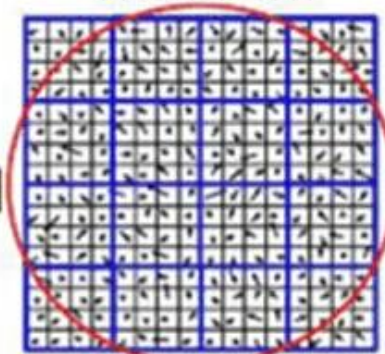
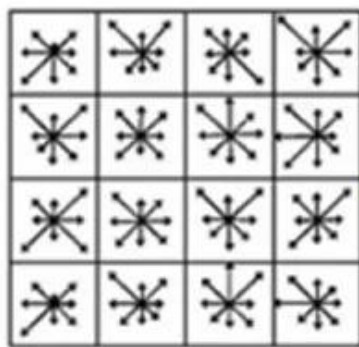


Compute the Difference of Gaussian Scale Space

For each pixel in DoG image, check if it is a stable keypoint or not



For each Keypoint, compute the gradient magnitude and orientation



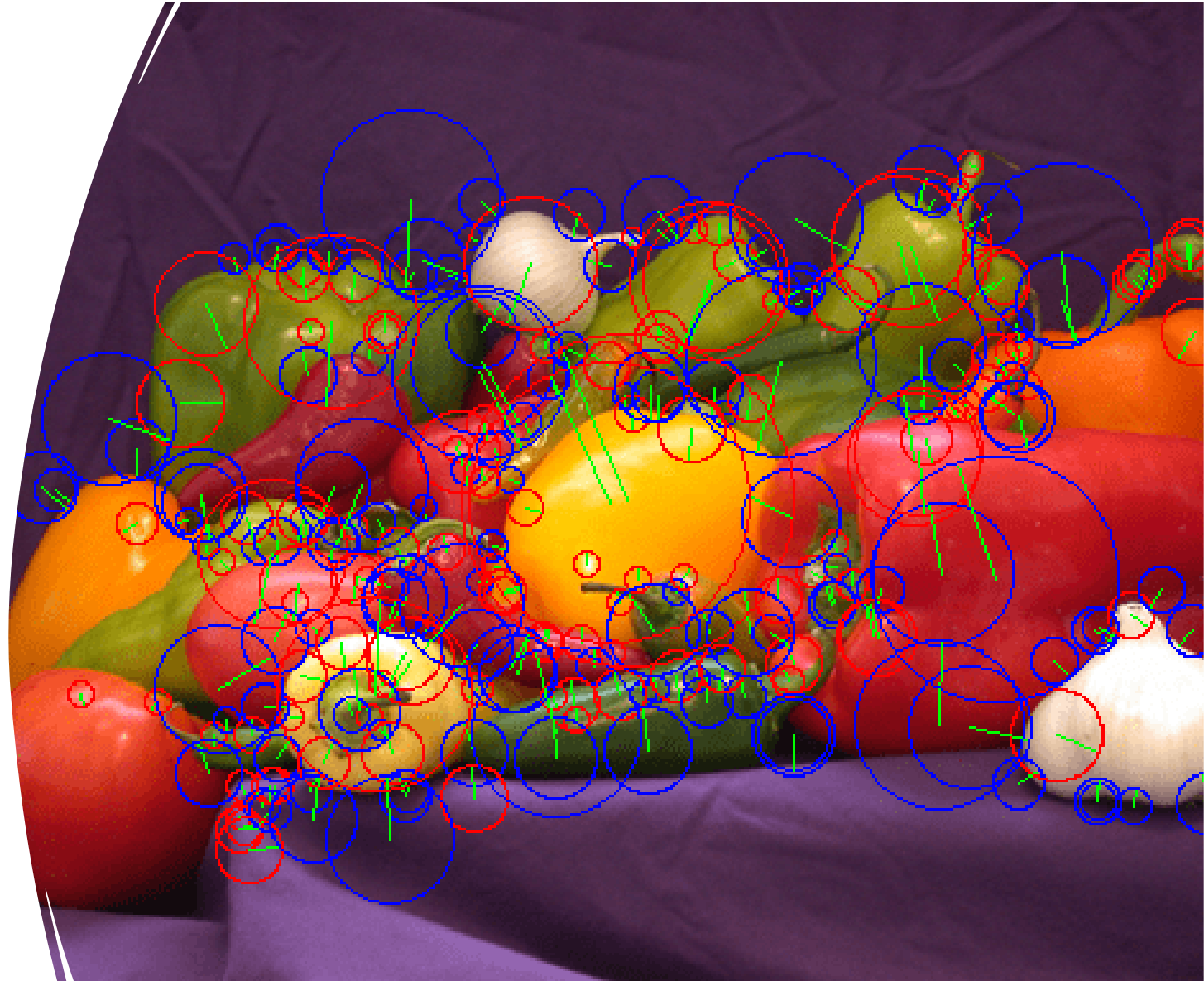
Histogram the Gradient values to build a 128 SIFT feature vector

128 Feature Vector

0.294	0.294	0.007	0.02	0.294	0.294	0.038	0.016	0.294	0.294	0.015	0.005	0.294	0.2
-------	-------	-------	------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-----

Speeded-Up Robust Features (SURF)

- Algoritmo desarrollado como una alternativa más rápida a SIFT, manteniendo una robustez similar.
- Mantiene los mismos pasos que SIFT pero aplica distintos enfoques para disminuir el tiempo.



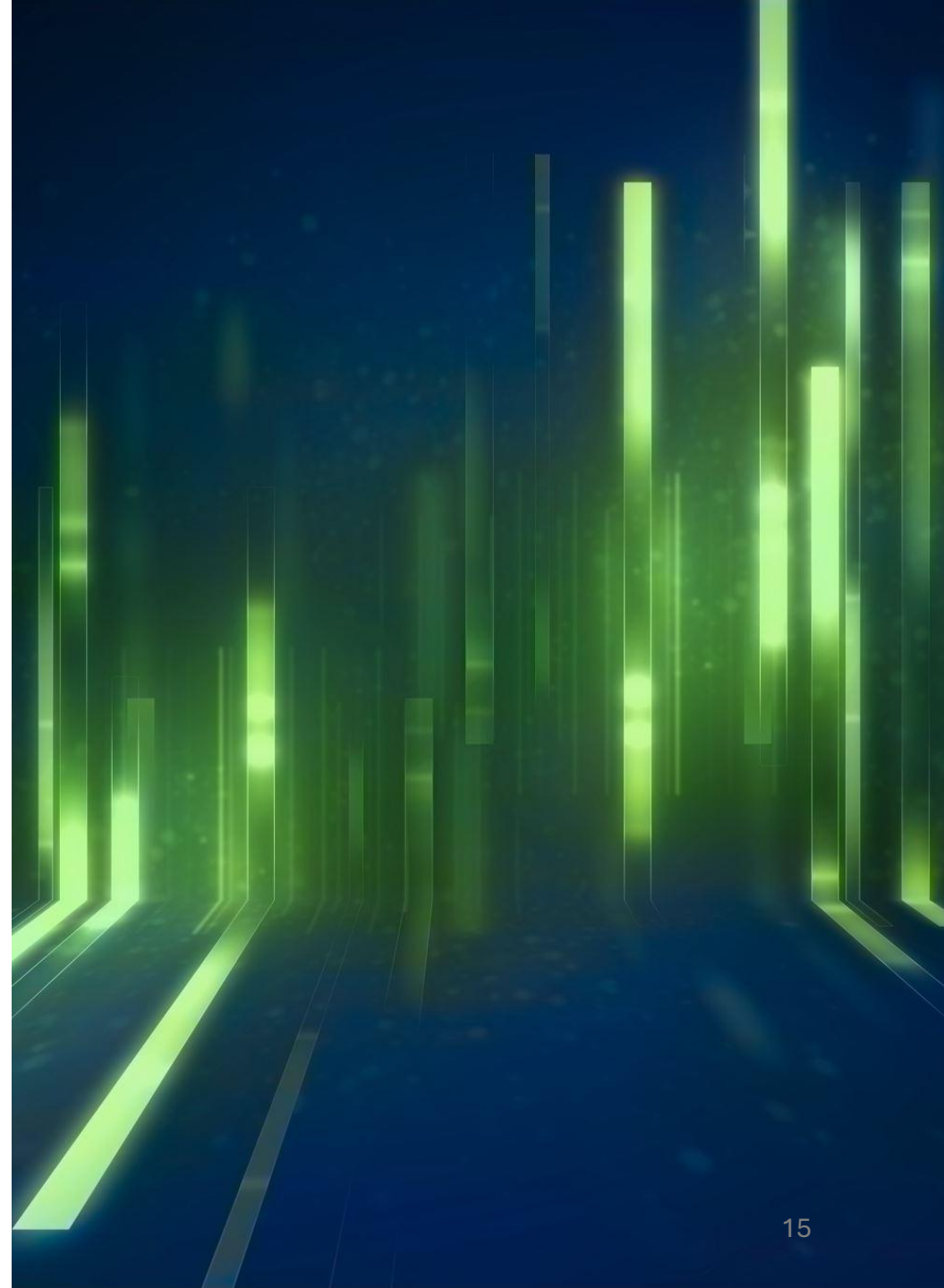
Estos algoritmos particularmente útiles para:

1. Reconocimiento de objetos
2. Seguimiento de objetos
3. Reconocimiento facial
4. Realidad aumentada.



Descriptores para audio

- Características temporales
- Características frecuenciales
- Características tiempo-frecuencia





Características temporales

- Energía/RMS

Mide la intensidad o potencia de la señal de audio en una ventana temporal.

$$E = \sum_{n=0}^{N-1} x[n]^2$$

$$RMS = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x[n]^2}$$

Energía

Si tengo una ventana con 4 muestras:

[0.5, -0.3, 0.8, -0.2]

$$E = (1/4) \times [(0.5)^2 + (-0.3)^2 + (0.8)^2 + (-0.2)^2]$$

$$E = (1/4) \times [0.25 + 0.09 + 0.64 + 0.04]$$

$$E = (1/4) \times 1.02$$

$$E = 0.255$$

- Zero Crossing Rate (ZCR)

Cuántas veces la señal cambia de signo.

$$ZCR = \frac{1}{2N} \sum_{n=1}^N |\text{sign}(x[n]) - \text{sign}(x[n-1])|$$

- Alto ZCR: sonido ruidoso
- Bajo ZCR: sonido armónico
- Voz vs Ruido
- Sonidos Percusivos

ZRC

Ventana: [0.5, 0.3, -0.2, -0.4, 0.1]

(+) (+) (-) (-) (+)

Comparaciones:

n=1: $\text{sign}(0.3) - \text{sign}(0.5) = 1 - 1 = 0 \rightarrow |0| = 0$

n=2: $\text{sign}(-0.2) - \text{sign}(0.3) = -1 - 1 = -2 \rightarrow |-2| = 2 \checkmark$ cruce

n=3: $\text{sign}(-0.4) - \text{sign}(-0.2) = -1 - (-1) = 0 \rightarrow |0| = 0$

n=4: $\text{sign}(0.1) - \text{sign}(-0.4) = 1 - (-1) = 2 \rightarrow |2| = 2 \checkmark$ cruce

Suma = $0 + 2 + 0 + 2 = 4$

ZCR = $(1/(2 \times 5)) \times 4 = 4/10 = 0.4$

- Autocorrelación

Mide qué tanto se parece una señal a una versión desplazada (retrasada o adelantada) de sí misma.

$$R[k] = \sum_{n=0}^{N-k-1} x[n] \cdot x[n-k]$$

Características Espectrales

- Estas se basan en la transformada de Fourier:

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi kn/N}$$

- **Spectral Centroid**

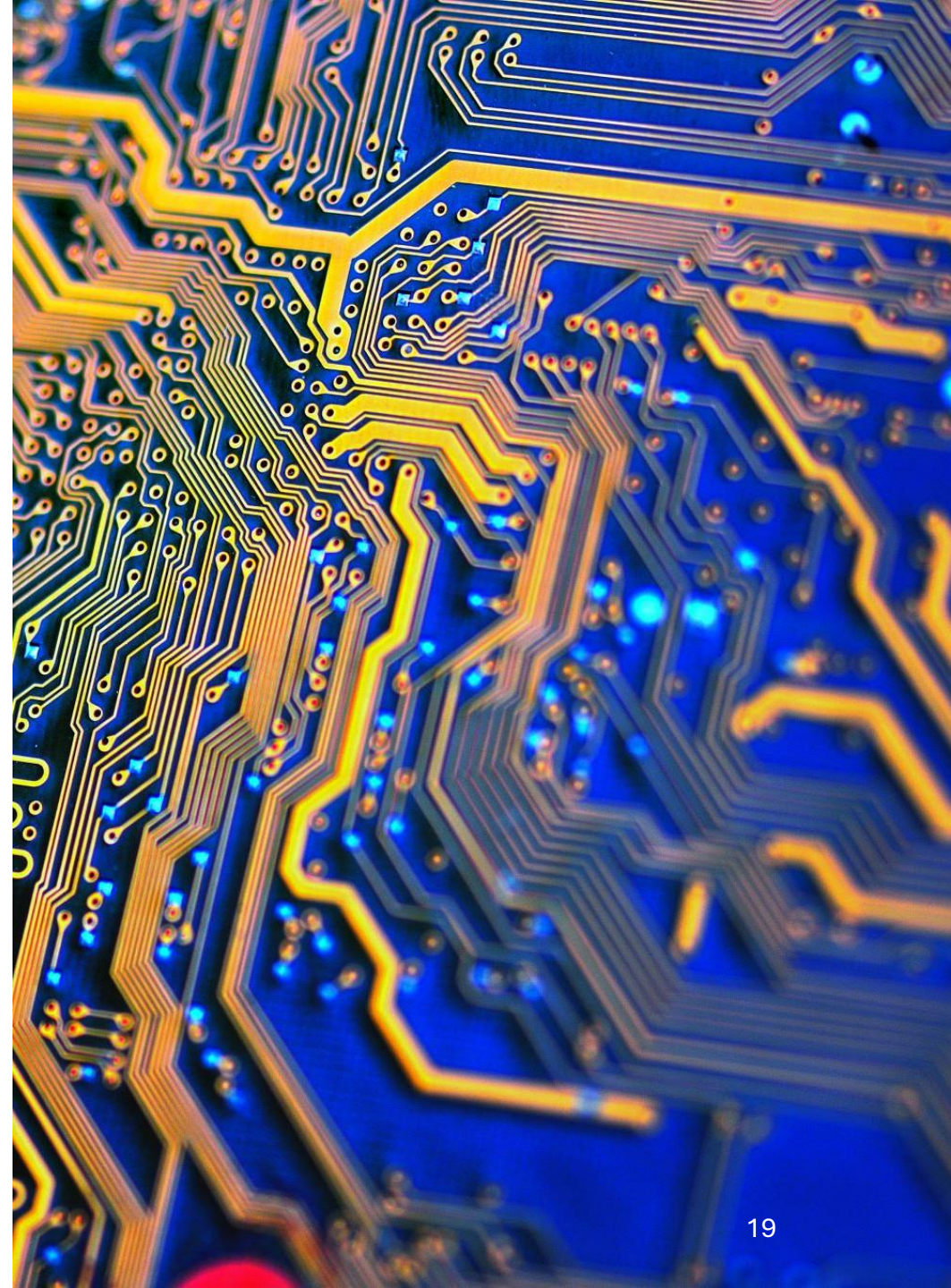
Centro de masa del espectro.

$$C = \frac{\sum_k f_k |X[k]|}{\sum_k |X[k]|}$$

$f[k]: k(sr/N)$

sr: frecuencia de muestreo

N: tamaño de la FFT



Para una señal:

$$f = [100\text{Hz}, 500\text{Hz}, 1000\text{Hz}, 2000\text{Hz}]$$

$$|X| = [0.1, 0.5, 0.3, 0.1]$$

$$\begin{aligned}\text{Numerador} &= 100 \times 0.1 + 500 \times 0.5 + 1000 \times 0.3 + 2000 \times 0.1 \\ &= 10 + 250 + 300 + 200 = 760\end{aligned}$$

$$\text{Denominador} = 0.1 + 0.5 + 0.3 + 0.1 = 1.0$$

$$\text{SC} = 760 / 1.0 = 760 \text{ Hz}$$

- Spectral Bandwidth

Mide dispersión espectral.

$$BW = \sqrt{\frac{\sum_k (f_k - C)^2 |X[k]|}{\sum_k |X[k]|}}$$

- Spectral Rolloff

Indica hasta qué frecuencia se concentra la energía.

$$\sum_{k=0}^R |X[k]| = \alpha \sum_{k=0}^K |X[k]|$$

α : *Umbral* (0.85 o 0.95)

Bandwidth

Supongamos $SC = 760\text{Hz}$

$f = [100, 500, 1000, 2000]$

$|X| = [0.1, 0.5, 0.3, 0.1]$

Desviaciones cuadradas:

$$(100-760)^2 = 435,600$$

$$(500-760)^2 = 67,600$$

$$(1000-760)^2 = 57,600$$

$$(2000-760)^2 = 1,537,600$$

$$\begin{aligned}\text{Numerador} &= 435600 \times 0.1 + 67600 \times 0.5 + 57600 \times 0.3 \\ &+ 1537600 \times 0.1\end{aligned}$$

$$= 43560 + 33800 + 17280 + 153760 = 248,400$$

$$\text{Denominador} = 0.1 + 0.5 + 0.3 + 0.1 = 1.0$$

$$BW = \sqrt{(248400 / 1.0)} = \sqrt{248400} \approx 498.4 \text{ Hz}$$

Rolloff

Espectro simplificado:

Bin: 0 1 2 3 4

$|X|$: 0.2 0.3 0.4 0.2 0.1

Suma total = 1.2

Buscamos donde se acumula el 85%:

$$85\% \text{ de } 1.2 = 1.02$$

Acumulado:

Hasta bin 0: 0.2

Hasta bin 1: $0.2 + 0.3 = 0.5$

Hasta bin 2: $0.5 + 0.4 = 0.9$

Hasta bin 3: $0.9 + 0.2 = 1.1 \checkmark$ (supera 1.02)

Rolloff bin = 3

Si $f[3] = 1500\text{Hz} \rightarrow \text{Rolloff} = 1500\text{Hz}$

- Spectral Flux

Mide cambio entre frames.

$$Flux = \sum_k (|X_t[k]| - |X_{t-1}[k]|)^2$$

- Spectral Flatness

Mide la uniformidad del espectro.

$$Flatness = \frac{Media\ Geometrica}{Media\ Aritmetica}$$

$$= \frac{(\prod_{k=1}^K |X[k]|)^{1/K}}{\frac{1}{K} \sum_k |X[k]|}$$

Flux

Frame anterior (t-1):

$|X_{t-1}| = [0.2, 0.5, 0.3]$

Frame actual (t):

$|X_t| = [0.3, 0.4, 0.8]$

Diferencias:

Bin 0: $0.3 - 0.2 = 0.1 \rightarrow (0.1)^2 = 0.01$

Bin 1: $0.4 - 0.5 = -0.1 \rightarrow (-0.1)^2 = 0.01$

Bin 2: $0.8 - 0.3 = 0.5 \rightarrow (0.5)^2 = 0.25$

$SF = 0.01 + 0.01 + 0.25 = 0.27$

Flatness

Espectro: $|X| = [0.1, 0.2, 0.15]$ (K=3)

Media geométrica:

Producto = $0.1 \times 0.2 \times 0.15 = 0.003$

Raíz cúbica = $0.003^{(1/3)} \approx 0.144$

Media aritmética:

$(1/3) \times (0.1 + 0.2 + 0.15) = (1/3) \times 0.45 = 0.15$

$SFlatness = 0.144 / 0.15 \approx 0.96$

(Valor alto $\rightarrow$ espectro relativamente plano)

Características tiempo-frecuencia

- MFCC – Mel Frequency Cepstral Coefficients

Son coeficientes que representan la envolvente espectral de una señal de audio en una escala perceptual similar a como el oído humano percibe el sonido.

Dan énfasis a características que son importantes para la percepción del habla humana mientras se descarta información menos importante.

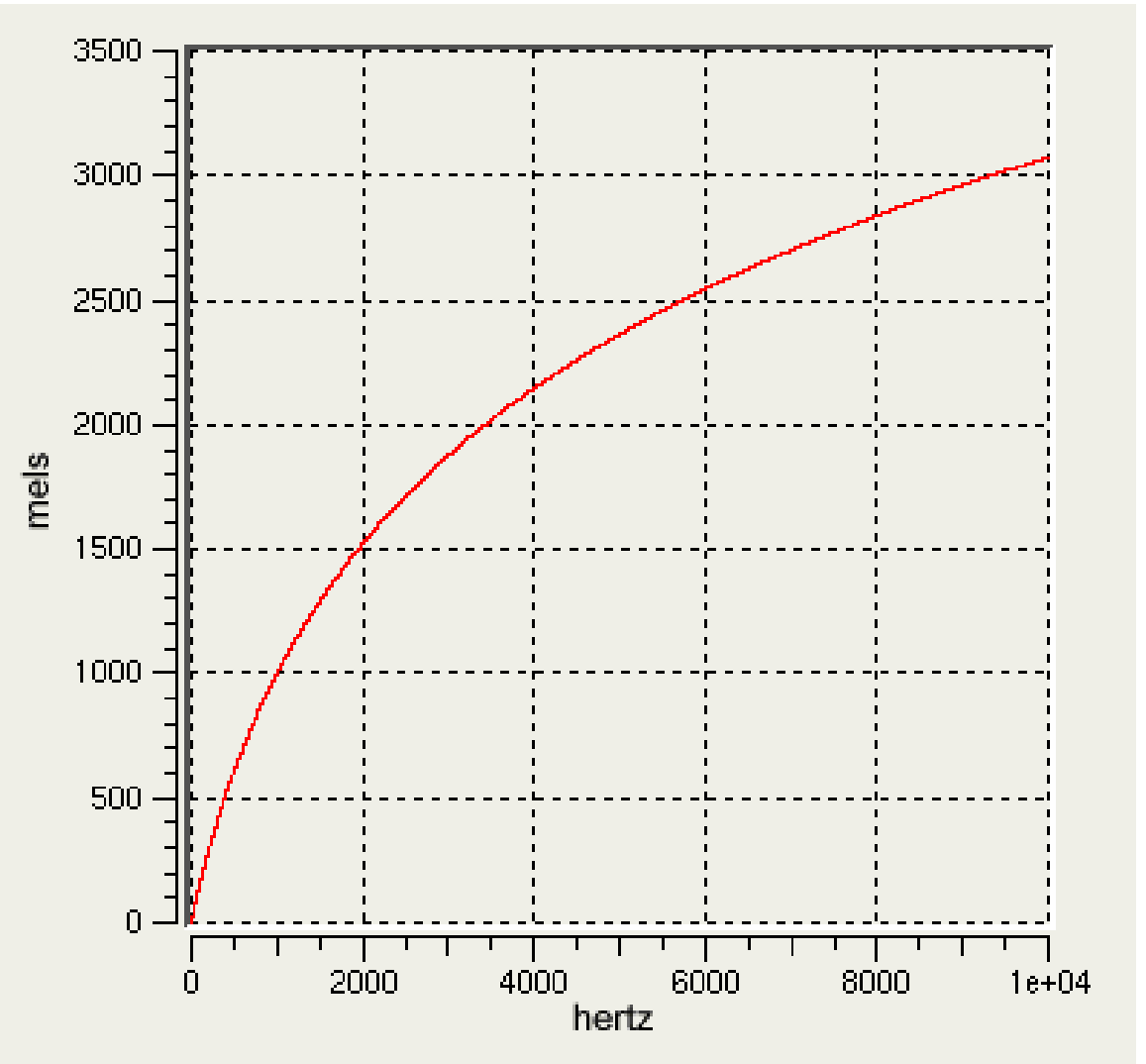


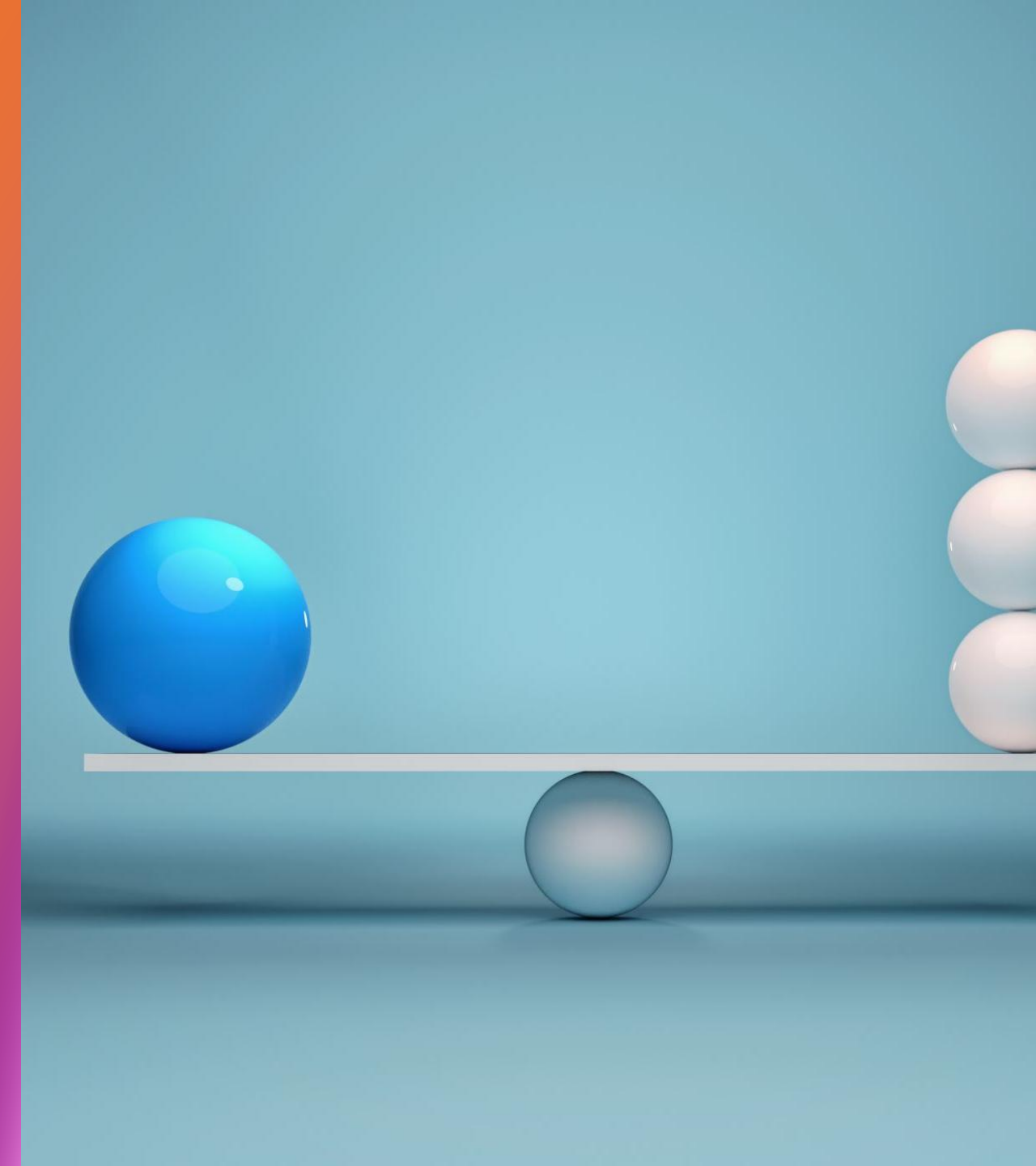
Escala MEL

- Es una escala diseñada para tratar de imitar la forma en que el ser humano percibe el audio. Es decir, a bajas frecuencias es aproximadamente lineal, a altas frecuencias se vuelve logarítmica.

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

$$f = 700 \cdot (10^{\frac{m}{2595}} - 1)$$



A blue sphere is positioned on the left side of a horizontal lever. On the right side of the lever, three white spheres are stacked vertically. The lever is balanced on a central pivot point, which is a small grey sphere. The background is a solid light blue color.

Consta de las siguientes etapas

1. Pre-énfasis

Se aplica un filtro paso altas para amplificar las altas frecuencias:

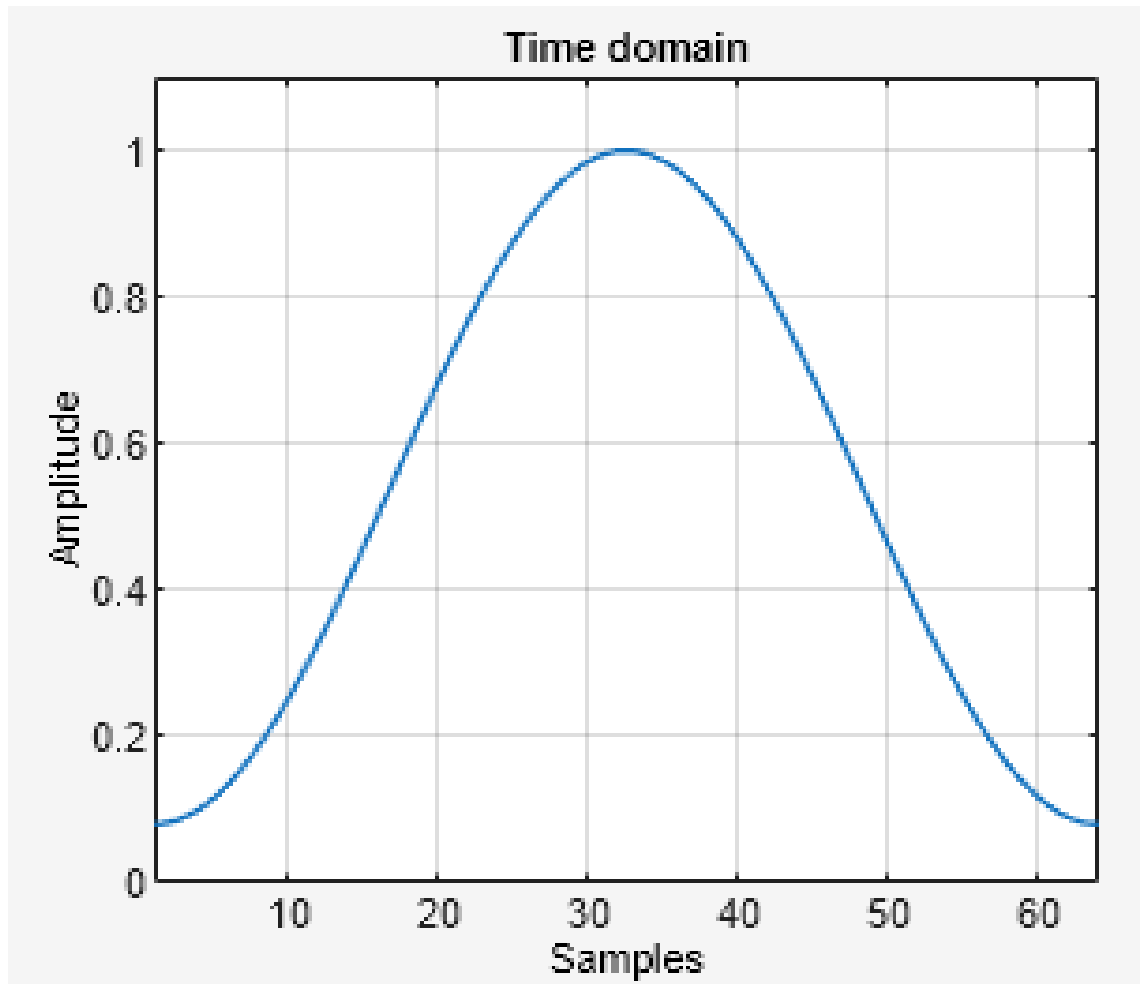
$$y[n] = x[n] - ax[n - 1], \quad a \approx 0.95$$

2. Framing

Se divide en ventanas locales:

20-40ms

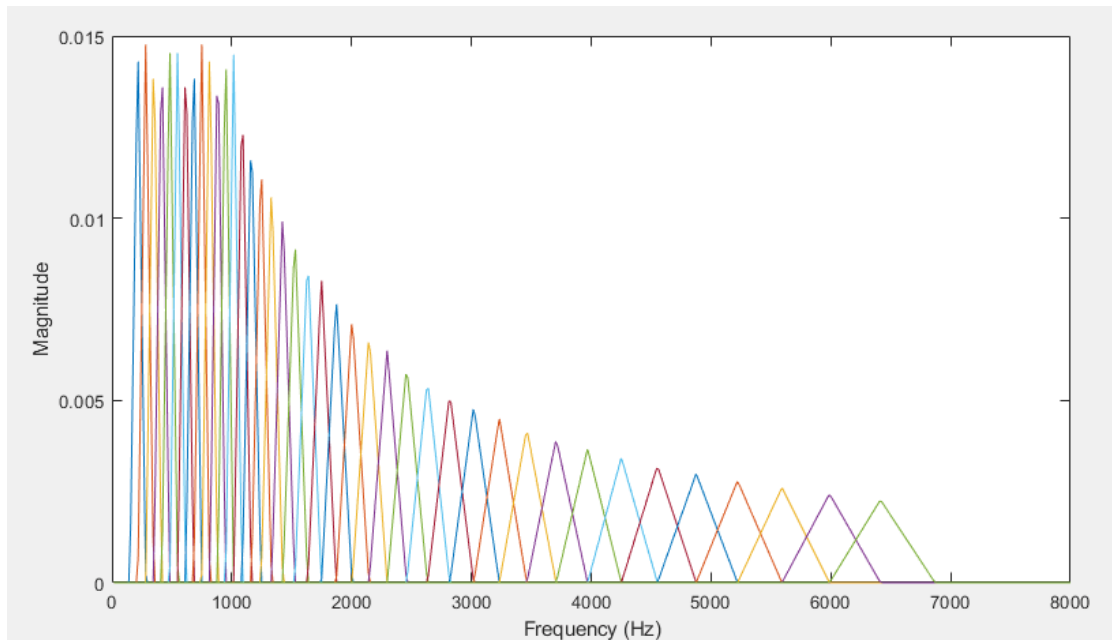
50% de solapamiento



3. Aplicación de una ventana.

Se aplica una ventana, por lo regular la de Hamming, esto para reducir los efectos de distorsión al aplicar la transformada de Fourier.

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$$



Cada filtro concentra la energía de esa banda en un solo valor:

$$energia[i] = \sum P[k]H_i[k]$$

Dando como resultado de 20 a 40 valores escalares.

4. Transformada de Fourier.

Se transforma cada ventana al dominio de la frecuencia y se obtiene el espectro de potencia.

$$X[k] = FFT(x[n]), \quad |X[k]|^2$$

5. Banco de Filtros Mel

Se aplica un conjunto de filtros triangulares espaciados según la escala Mel. Se utilizan de 20 a 40 filtros Mel.

6. Logaritmo.

Se aplica logaritmo a la energía de cada filtro.

$$\text{Log}(E_m)$$

7. Transformada Coseno Discreta (DCT)

Se emplea la DCT-II:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cos \left[\frac{\pi}{N} \left(n + \frac{1}{2} \right) k \right]$$

Me entrega un vector de coeficientes reales:
[X[0], X[1],..., X[N-1]].

Para este caso, de la misma longitud que
nuestro numero de filtros Mel.

Pero por lo regular se utilizan los primeros
12-13 coeficientes. **Estos son los MFCCs.**

- Esto se aplica para cada ventana que generada de tal forma que obtenemos una matriz de la forma:

	c1	c2	c3	...	c13
v1	x	x	x	...	x
v2	x	x	x	...	x
v3	x	x	x	...	x
...	...	...	...	...	...
vt	x	x	x	...	x

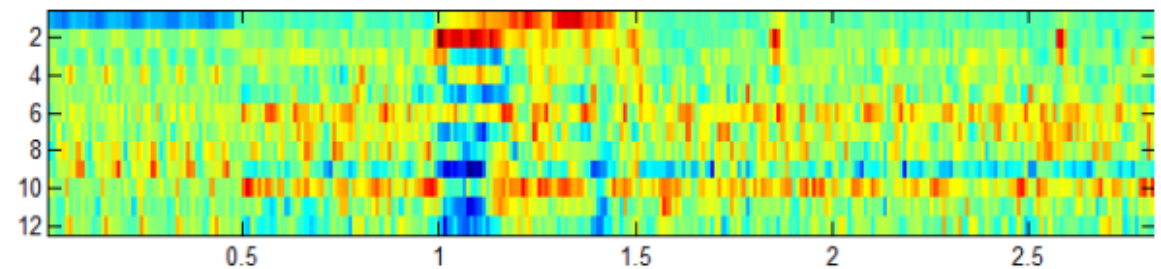
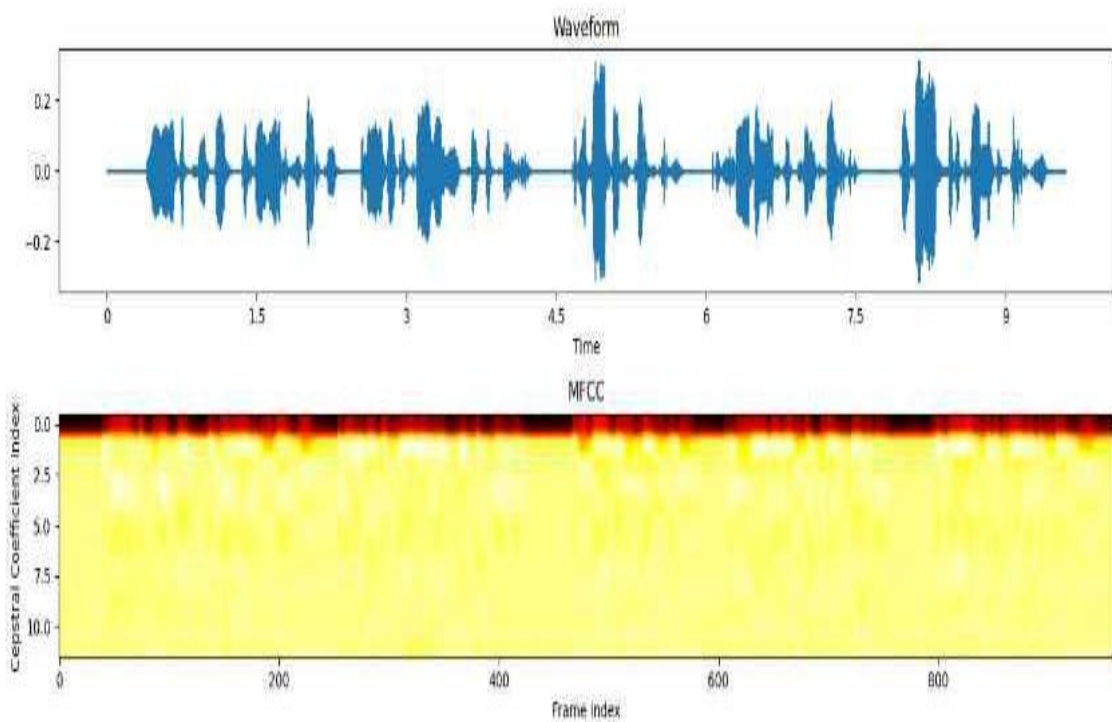


Fig.11. Plot of Mel frequency Cepstrum

Delta y Delta-Delta

- Utilizadas para modelar la dinámica temporal de la señal. Se calculan a partir de las 12 o 13 MFCCs, y generan la misma cantidad, es decir: 13 MFCCs, 13 deltas, 13 delta-delta

$$\Delta c_t = \frac{\sum_{n=1}^N n (c_{t+n} - c_{t-n})}{2 \sum_{n=1}^N n^2} \quad \Delta^2 c_t = \frac{\sum_{n=1}^N n (\Delta c_{t+n} - \Delta c_{t-n})}{2 \sum_{n=1}^N n^2}$$

De esta forma, se tendría una matriz de #ventanas x 39

¿Cómo obtener un vector de características?

- Para cada característica extraída por ventana, calcular:
 - **Media** (μ)
 - **Desviación estándar** (σ)
 - **Máximo**
 - **Mínimo**
 - Mediana, percentiles, rango, etc.

Si se extrae 13 MFCCs por frame:

- 13 medias + 13 std + 13 max + 13 min = 52 características

Si se añade 2 temporales (energía, ZCR):

- $2 \times 4 = 8$ características adicionales

Si se añade 5 espectrales:

- $5 \times 4 = 20$ características adicionales

Total: $52 + 8 + 20 = 80$ dimensiones por audio

Para la próxima clase...

- Selección de características

