Mel-Frequency Cepstral Coefficient

Diana Patricia Tobón Vallejo, PhD

Tratamiento de Señales III
Facultad de Ingeniería
Universidad de Antioquia

2024

Material elaborado por: Hernán Felipe García Arias

Definición

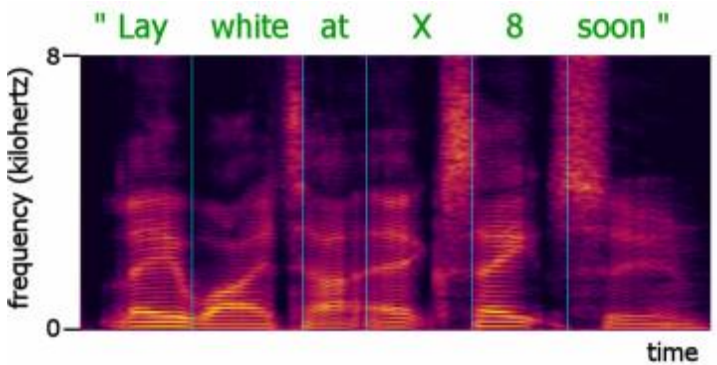
- Mel Frequency Cepstral Coefficients (MFCC) es una forma de extraer características de un audio.
- El MFCC usa la escala MEL para dividir la banda de frecuencia en sub-bandas y luego extrae los coeficientes cepstrales usando la transformada discreta del coseno (DCT).
- La escala MEL se basa en la forma en que los humanos distinguen entre frecuencias, lo que hace que sea muy conveniente procesar sonidos.

Percepción del sonido de la voz humana

- El rango de frecuencia fundamental de la voz de los seres humanos adultos está entre 85Hz y 255Hz (85Hz a 180Hz para hombres y 165Hz a 255Hz para mujeres).

- Además de la frecuencia fundamental, hay armónicos de frecuencias fundamentales.[1]

- Si, por ejemplo, la frecuencia fundamental es 100 Hz, entonces su segundo armónico será 200 Hz, el tercer armónico es 300 Hz y así sucesivamente.



El color representa la potencia de frecuencia en ese punto (amarillo más fuerte y negro más débil)

[1] Los armónicos son multiplicaciones enteras de la frecuencia fundamental.

Escala de MEL

- Los seres humanos pueden oir aproximadamente entre $20Hz$ y $20kHz$. La percepción del sonido no es lineal [2] y puede distinguir mejor entre los sonidos de baja frecuencia que los de alta frecuencia.

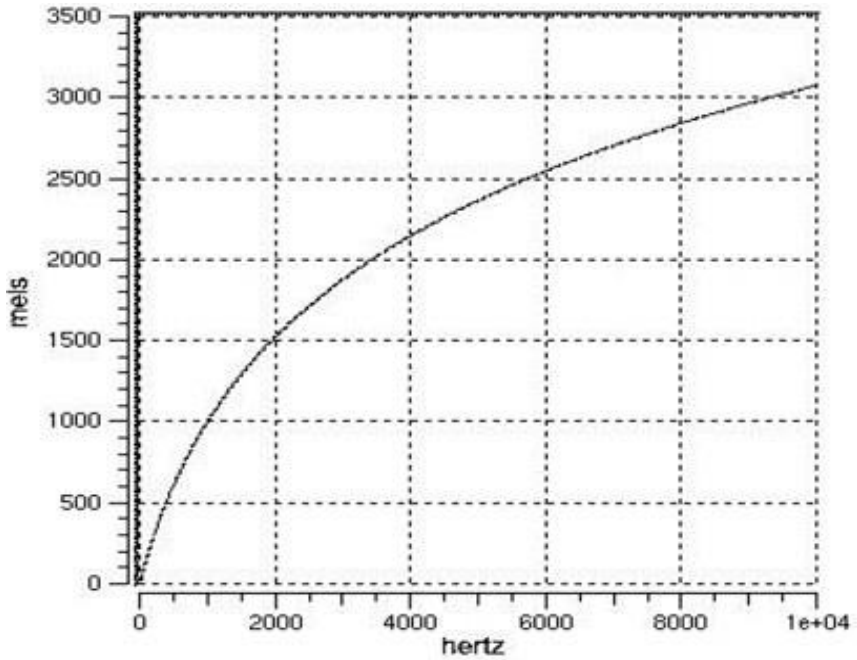
- Los humanos pueden escuchar claramente la diferencia entre $100Hz$ y $200Hz$, pero no entre $15kHz$ y $15.1kHz$.

- Una escala MEL es una unidad de PITCH propuesta por Stevens, Volkmann y Newmann en 1937.

Escala de MEL

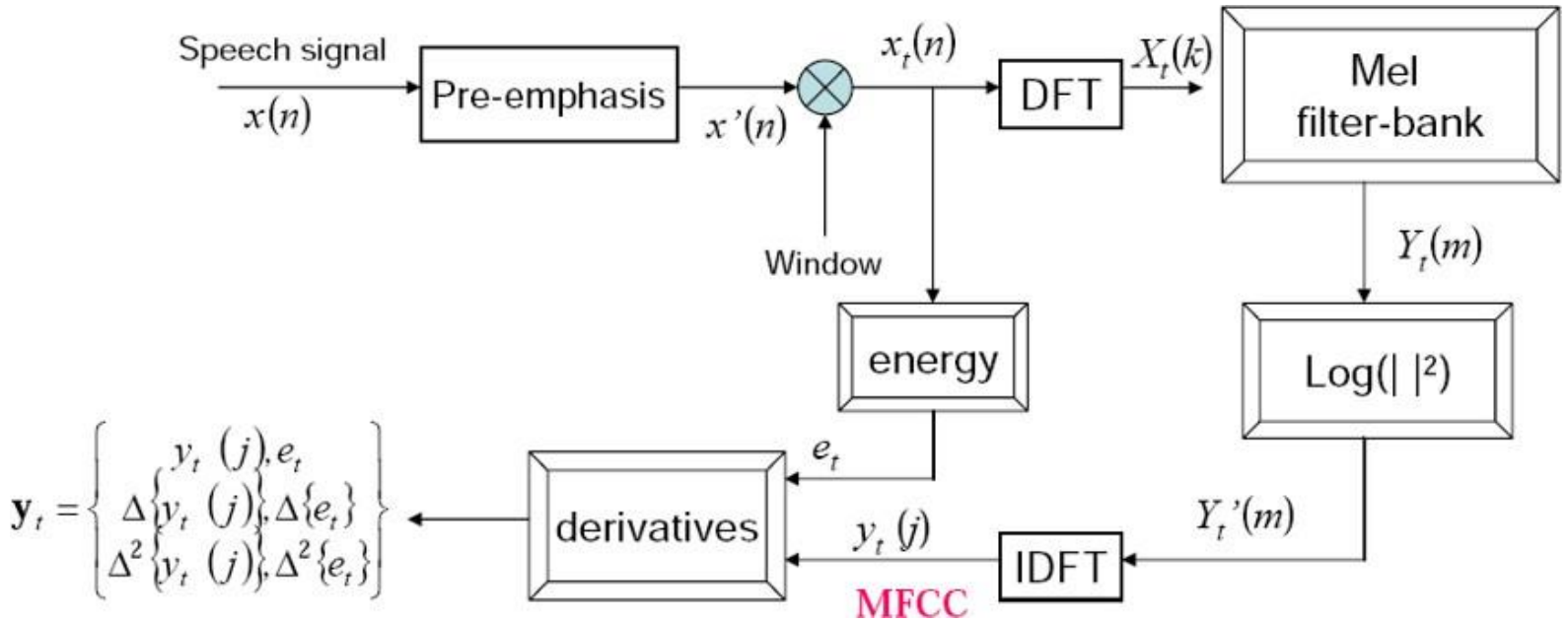
- La escala MEL es una escala de tonos que los oyentes consideran iguales en distancia entre sí [3] [4].

- Debido a como los humanos perciben el sonido, la escala MEL es una escala no lejana y las distancias entre los tonos aumentan con la frecuencia.



Coeficientes Cepstrales de Mel (MFCC)

- Representación espectral más utilizada en *automatic speech recognition* (ASR).

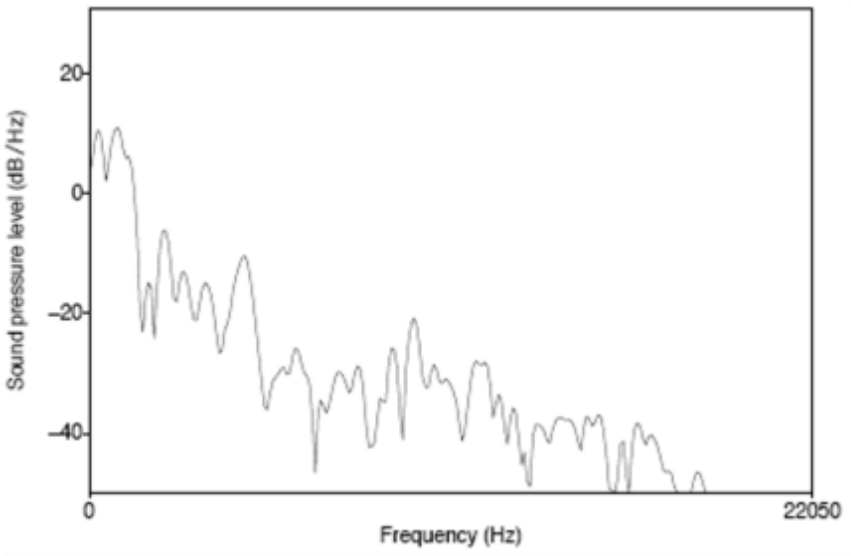


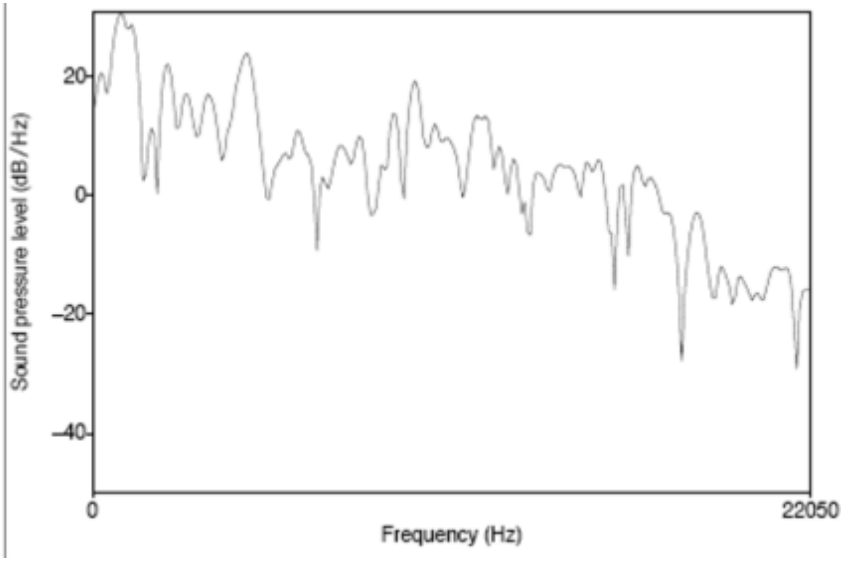
Pre-énfasis

- **Pre-énfasis:** aumentar la energía en las altas frecuencias
- **P:** ¿Por qué hacer esto?
- **R:** El espectro de los segmentos sonoros tiene más energía a frecuencias más bajas que a frecuencias más altas.
  - Esto se llama inclinación espectral.
  - La inclinación espectral es causada por la naturaleza del pulso glótico.
- El aumento de la energía de alta frecuencia brinda más información al modelo acústico.

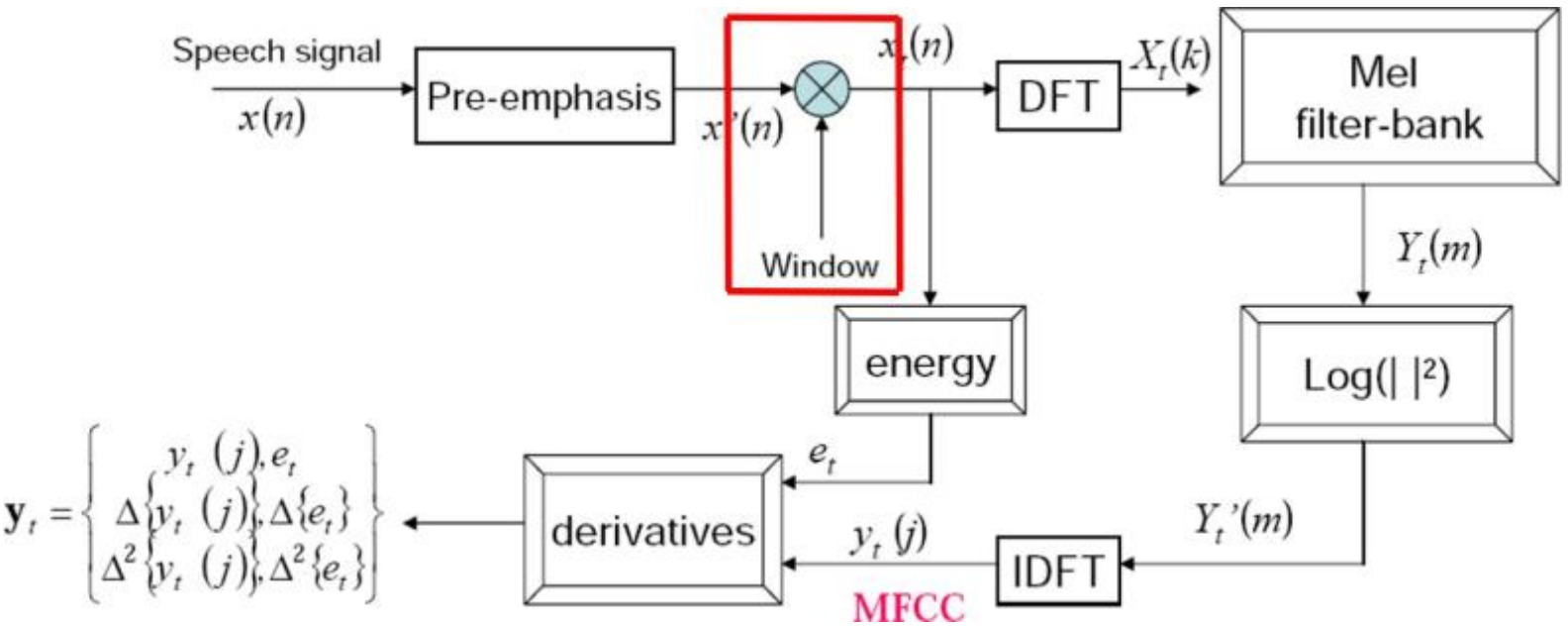
Ejemplo de pre-énfasis

- Antes y después del pre-énfasis
  - Slice espectral de la vocal [aa]

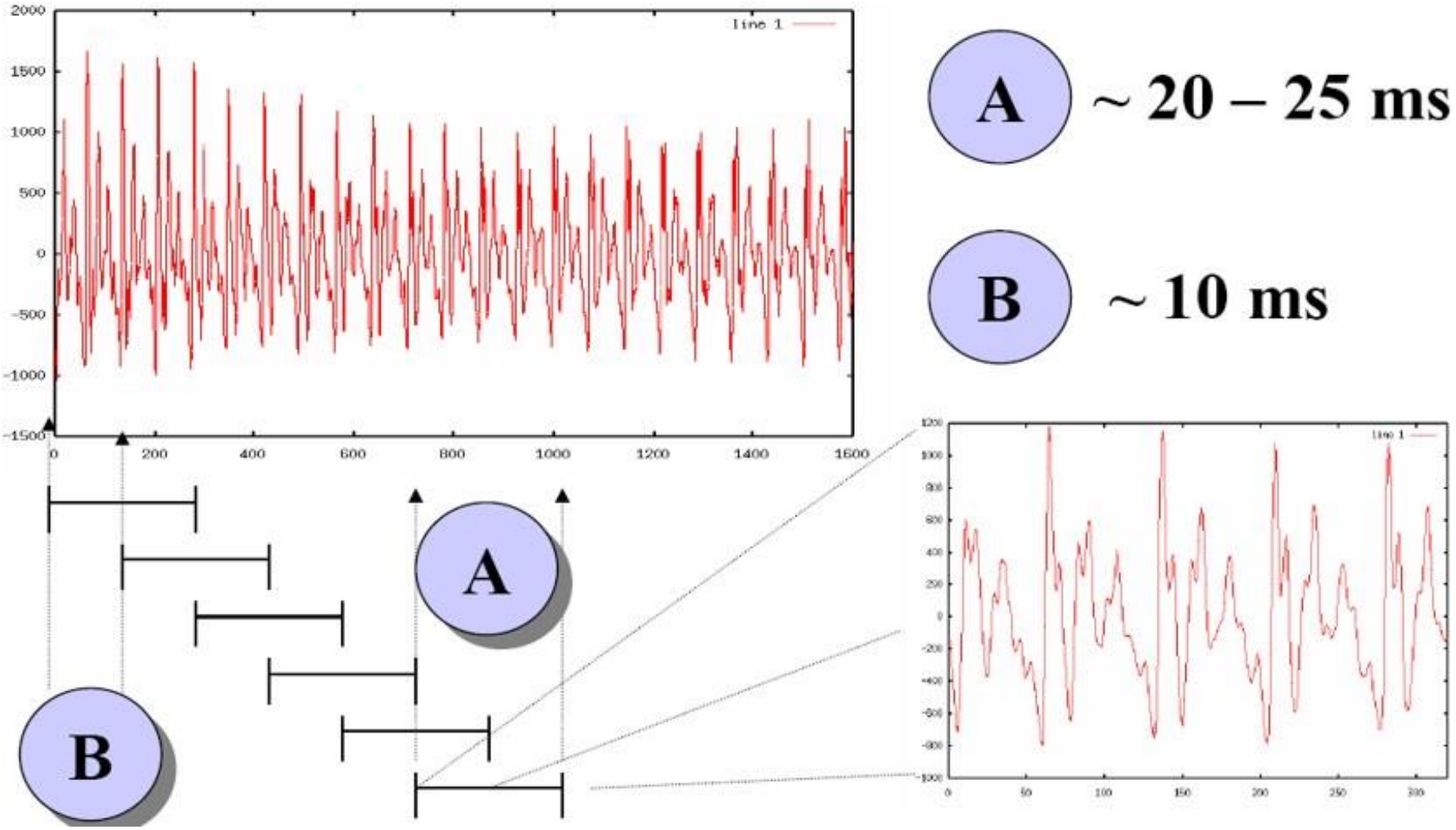




MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
$x_t(n)$
Window
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
MFCC
energy
e_t
derivatives
$\mathbf{y}_t = \begin{Bmatrix} y_t(j), e_t \\ \Delta\{y_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t(j)\}, \Delta^2\{e_t\} \end{Bmatrix}$

Enventanado (Windowing)

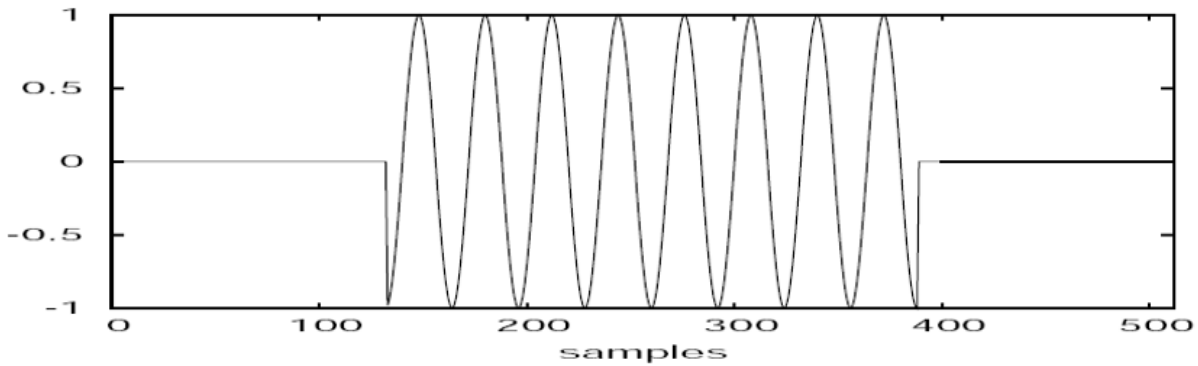


Enventanado (Windowing)

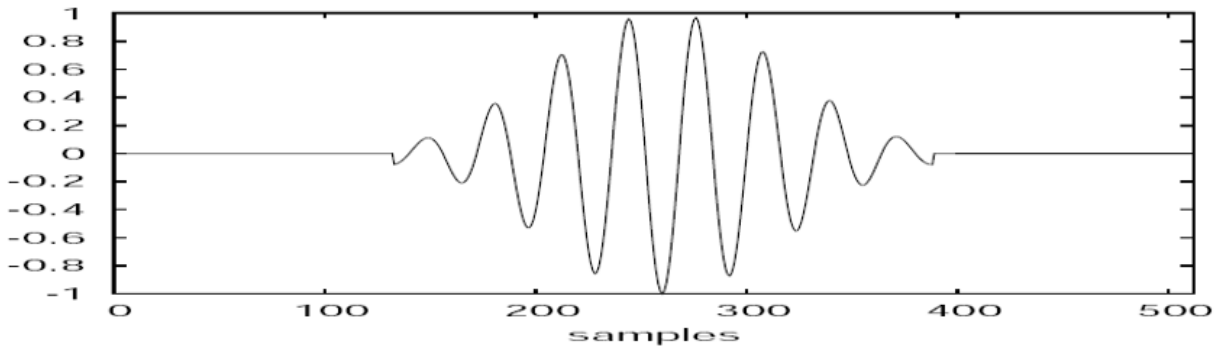
- ¿Por qué dividir la señal de voz en sucesivos marcos (*frames*) superpuestos?
  - El habla no es una señal estacionaria; queremos información sobre una región lo suficientemente pequeña como para que la información espectral sea una pista útil.

- Marcos (*frames*)
  - Tamaño del *frame*: normalmente, 20-25 ms
  - Desplazamiento del *frame*: el periodo de tiempo entre *frames* sucesivos, por lo general, 5-10 ms

Ventana en el dominio temporal

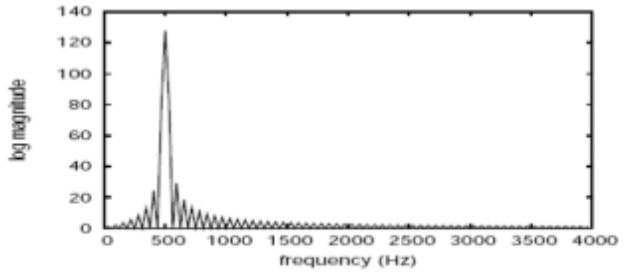


(a) Rectangular window

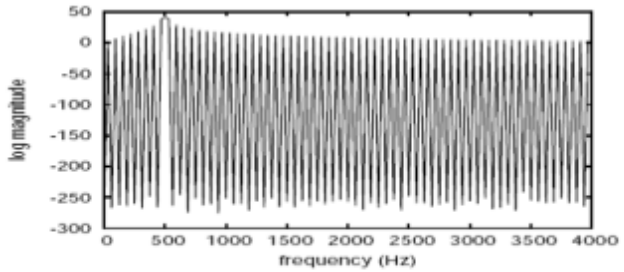


(c) Hamming window

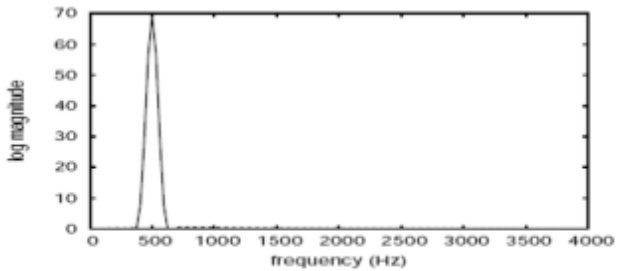
Ventana en el dominio de la frecuencia



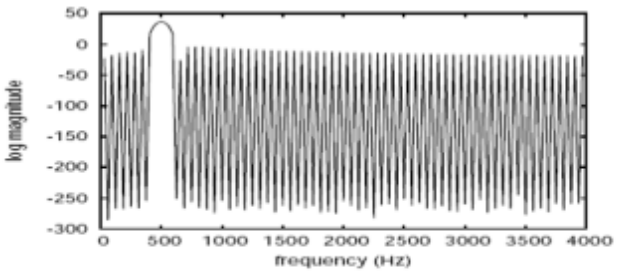
(a) Rectangular window



(b) Rectangular window

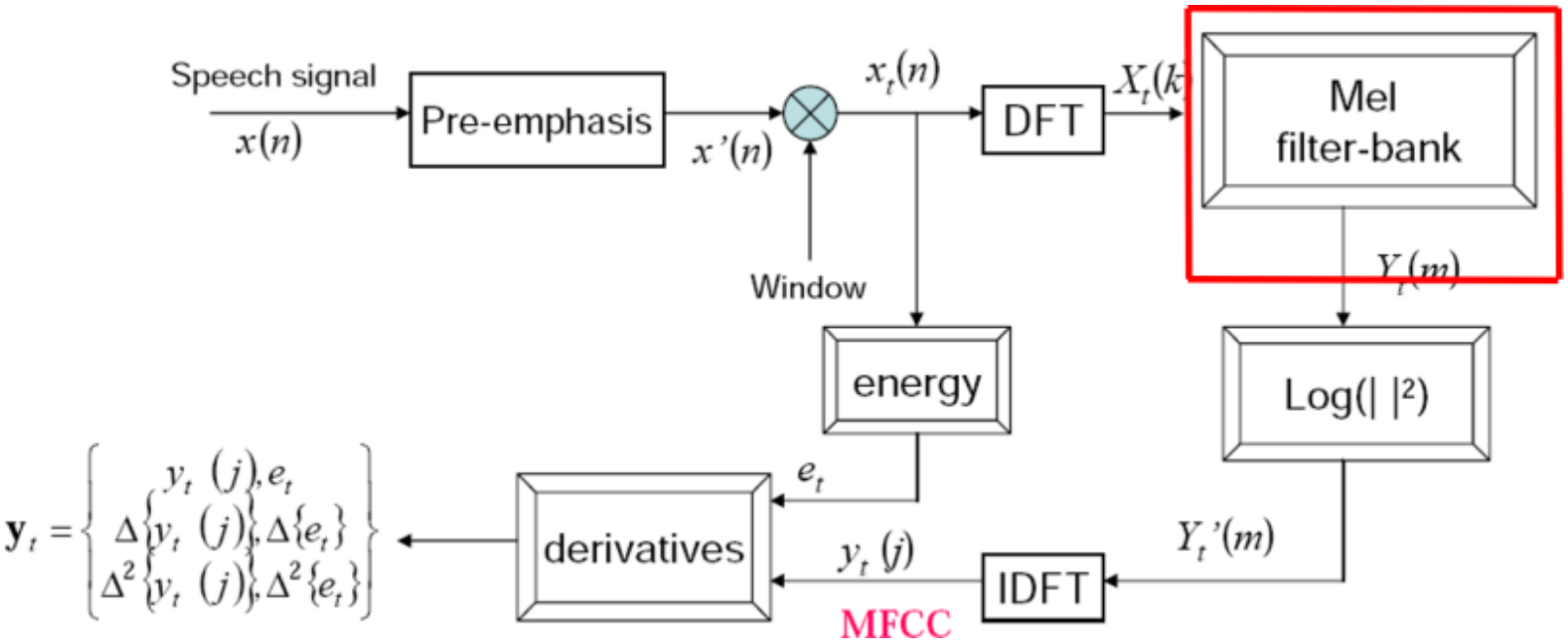


(e) Hamming window



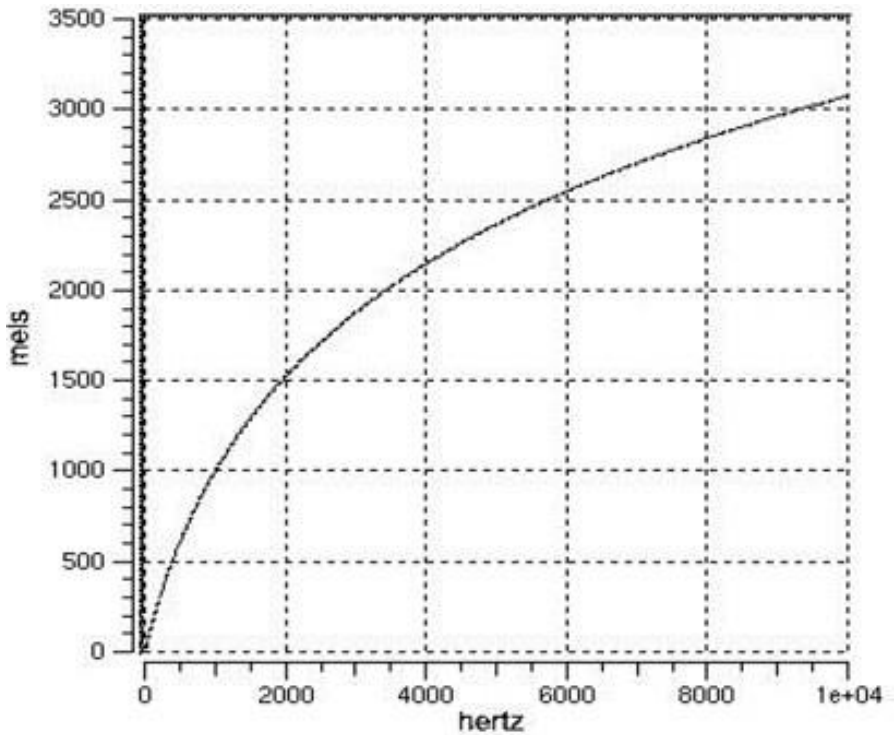
(f) Hamming window

MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
MFCC
energy
e_t
derivatives
$\mathbf{y}_t = \begin{Bmatrix} y_t(j), e_t \\ \Delta\{y_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t(j)\}, \Delta^2\{e_t\} \end{Bmatrix}$

Escala de MEL

- El oído humano no es igualmente sensible a todas las bandas de frecuencia. Menos sensible a frecuencias más altas, aproximadamente > 1000*Hz.*
- Es decir, la percepción humana de la frecuencia no es lineal.



Escala de MEL

- Un mel es una unidad de tono:
  - Definición:
    - Pares de sonidos perceptualmente equidistantes en el tono.
    - Están separados por un número igual de mels.

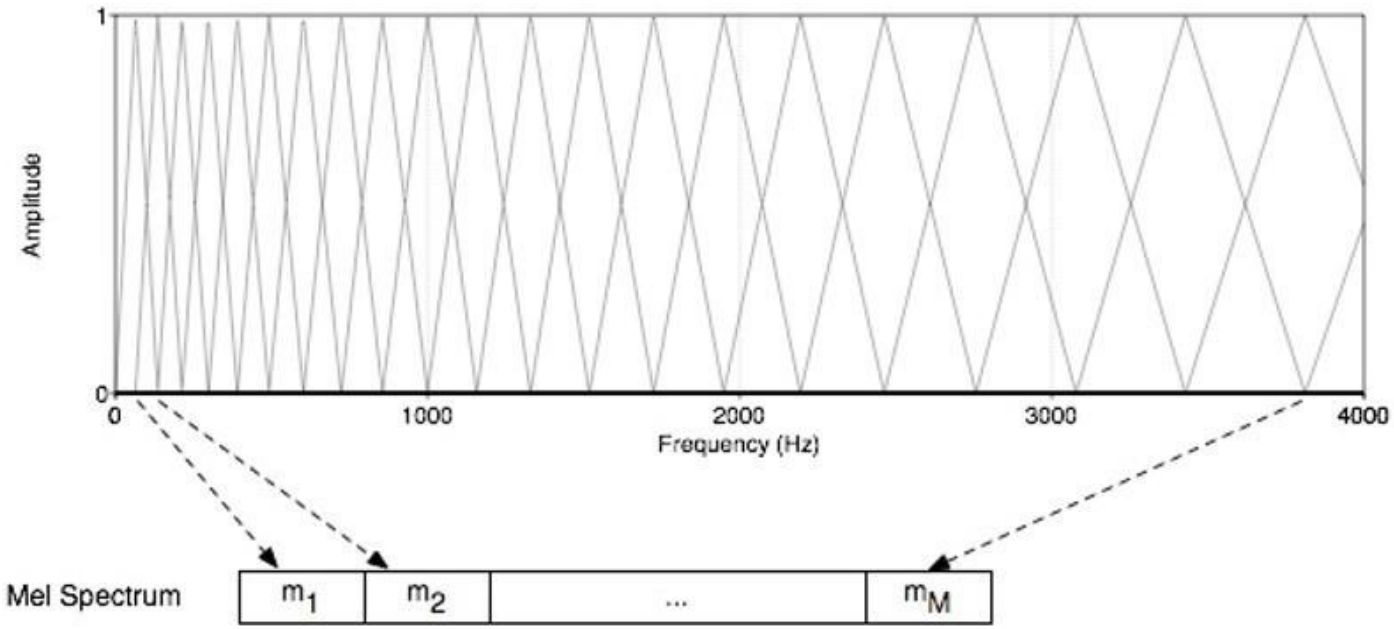
- La escala Mel es aproximadamente lineal por debajo de 1 kHz y logarítmica por encima de 1kHz.

- Definición:

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

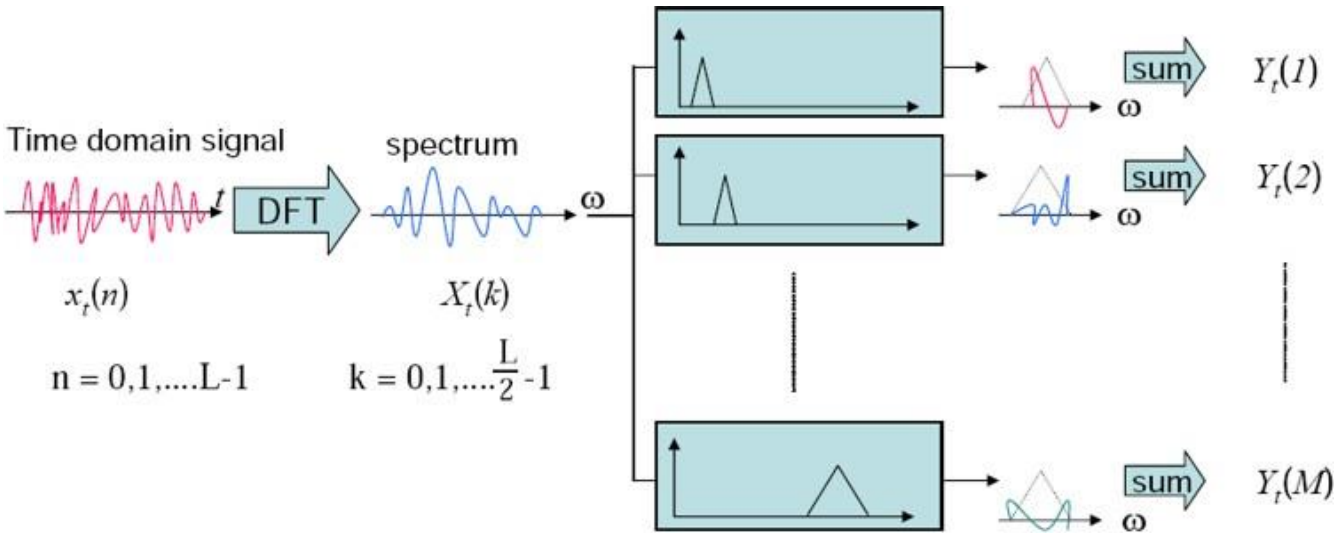
Procesamiento en el banco de filtros MEL

- Espaciado uniformemente antes de 1*kHz*
- Escala logarítmica después de 1*kHz*



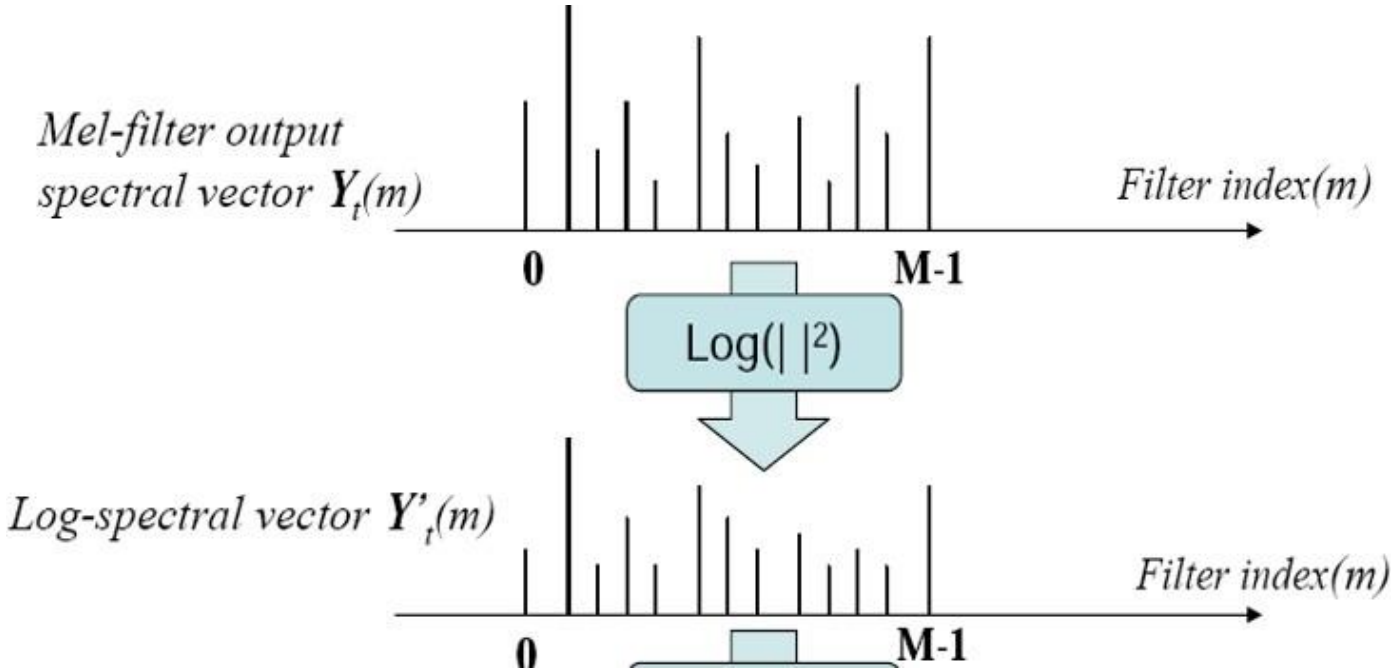
Procesamiento en el banco de filtros MEL

- Aplicar el banco de filtros según escala Mel al espectro.
- Cada salida de filtro es la suma de sus componentes espectrales filtrados.



Cálculo de la energía logarítmica

- Calcule el logaritmo de la magnitud cuadrada de la salida del banco de filtros Mel



Cálculo de la energía logarítmica

- **¿Por qué el logaritmo de la energía?**
  - El logaritmo comprime el rango dinámico de valores.
    - La respuesta humana al nivel de la señal es logarítmica.
    - Los seres humanos son menos sensibles a ligeras diferencias de amplitud en amplitudes altas que en amplitudes bajas.
  - Hace que las estimaciones de frecuencia sean menos sensibles a ligeras variaciones en la entrada (variación de potencia debido a que la boca del hablante se acerca al micrófono).
  - La información de fase no es útil en el habla.

Frecuencia Cepstral de Mel

- El cepstrum requiere análisis de Fourier [2]
- Pero vamos del espacio de frecuencias al tiempo
- Entonces, en realidad, aplicamos DFT inversa

$$y_t[k] = \sum_{m=1}^{M} \log\left(|Y_t(m)|\right) \cos\left(k(m-0.5)\frac{\pi}{M}\right), \quad k = 0, \ldots, J \qquad (1)$$

- **Hint!**: dado que el espectro de potencia logarítmica es real y simétrico, la DFT inversa se reduce a una Transformada de coseno discreta (DCT).

[2] El cepstrum de una señal es el resultado de calcular la transformada de Fourier inversa del espectro de la señal estudiada en escala logarítmica (dB). El cepstrum es complejo y, por tanto, tiene su parte real y su parte imaginaria.

Características típicas de los MFCC

- Window size: 25ms
- Window shift: 10ms
- Pre-emphasis coefficient: 0.97
- **MFCC**:
  - 1 energy feature MFCC[0]
  - 12 MFCC (mel frequency cepstral coefficients) MFCC[1]-MFCC[12]
  - 12 delta MFCC features MFCC[13]-MFCC[24]
  - 12 double-delta MFCC features MFCC[25]-MFCC[36]
  - 1 delta energy feature MFCC[37]
  - 1 double-delta energy feature MFCC[38]
- Total 39-dimensional features

¿Por qué los MFCC son tan populares?

- Eficiente para calcular
- Incorpora una escala de frecuencia Mel perceptual
- Separa la fuente y el filtro
- IDFT (DCT) decorrelaciona las características

https://colab.research.google.com/github/iiscleap/coswara-blog/blob/master/_notebooks/2020-08-20-mfcc.ipynb

Referencias

1 https://en.wikipedia.org/wiki/Voice_frequency

2 https://en.wikipedia.org/wiki/Hearing_range

3 https://www.sfu.ca/sonic-studio/handbook/Mel.html

4 https://en.wikipedia.org/wiki/Mel_scale

5 http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

6 https://en.wikipedia.org/wiki/Discrete_cosine_transform