MFCC Extraction Math

$S_1[n] \rightarrow$ signal

$S_1[n] \longrightarrow \boxed{\text{Pre-emphasis}} \longrightarrow \boxed{\text{Windowing}} \longrightarrow s[n]$

$s[n] \xrightarrow[(N > \text{Window size})]{\text{N-point DFT}} S(K) \longrightarrow |S(K)|$
$\downarrow$
Spectrum

$K = 0, 1, 2, \ldots, \frac{N}{2}$

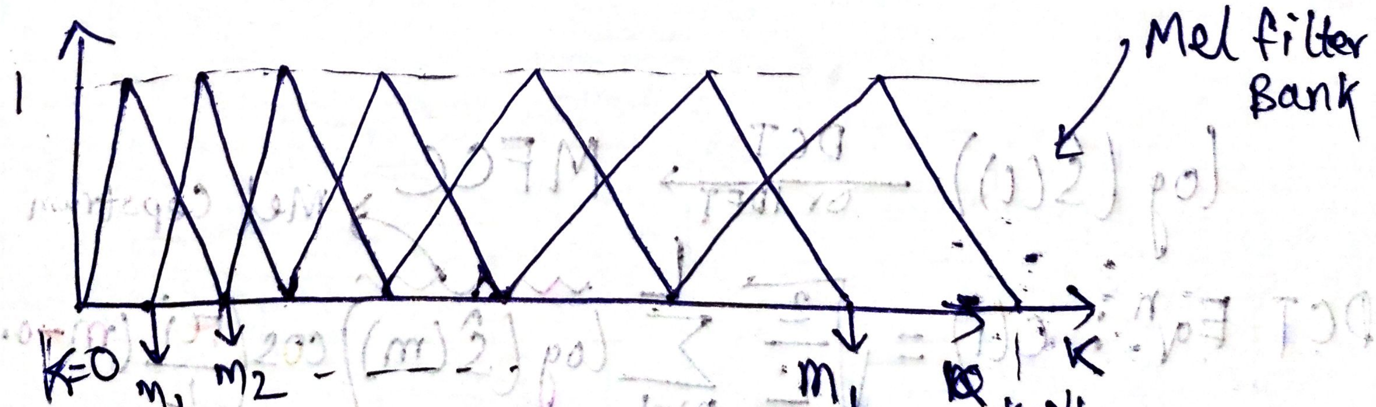
$|S(K)| \xrightarrow[\text{Bank}]{\text{Mel filter}} |\hat{S}(l)|$, l is the no. of Mel filters required

$$\hat{S}(l) = \sum_{K=0}^{N/2} S(K) \cdot M_l(K)$$

lth filter from filter Bank

$K \longleftrightarrow \frac{K f_s}{N}$ Hz ($f_s \rightarrow$ Sampling freq.)

$M_l(K) \rightarrow l$th filter weight



$m_1, m_2, \ldots, m_l$ are mel spectrum

Though Bandwidth is same in each filter, yet in higher frequency region, the difference from seems to high as Mel-scale is followed here —

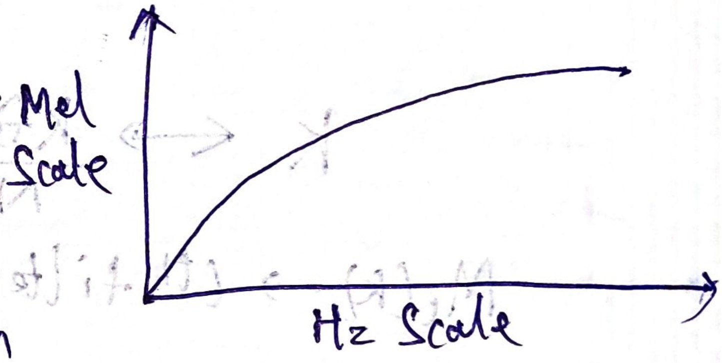
$$\boxed{f(\text{Mel}) = 2595 * \log_{10}\left(1 + \frac{f(\text{Hz})}{700}\right)}$$

Considering, l filters covers $k = 0, \ldots, N/2$

$$\hat{S}(l) = \sum_{k=0}^{N/2} S(k) \cdot M_l(k)$$

By following mel scale, energy of signal at higher frequencies will be increased, for which frequencies at that range could be easily perceived by humans.

$$\underset{\text{Mel Spectrum}}{\hat{S}(l)} \xrightarrow{\log} \underset{\text{Mel Cepstrum}}{\log(\hat{S}(l))}$$



$$\log(\hat{S}(l)) \xrightarrow[\text{or IDFT}]{\text{DCT}} \text{MFCC} \rightarrow \text{Mel Cepstrum}$$

DCT Eq^n : $c(i) = \sqrt{\frac{2}{L}} \sum_{m=1}^{L} \log(\hat{S}(m)) \cos\left(\frac{\pi i}{L}(m - 0.5)\right)$

C is no. of Cepstrum Coeff. Desired $\quad i = 0, 1, \ldots, C-1$