The mel-frequency cepstrum coefficients (MFCC) is a different method for recognition. The key difference with merely cepstrum coefficients (CC) is that it uses a mel scale rather than a hertz scale, which is adapted from human auditory perception system and is nonlinear in frequency. Mathematically, after achieving the CC of each frame, rather than finding the smoothing version, we pass it through an array of filters H (the properties of which is listed below). This is done by computing $Y = HWX$ where each column of X is the \texttt{NFFT}-point fft (where we can choose \texttt{NFFT} to be a power of 2 for faster computation) of each frame. $W\in\mathcal{R}^{M\times NFFT}$ is constructed to get rid of the aliasing artifact from log spectrum, as it has a $M\times M$ identity submatrix along with a $M\times M-2$ zero submatrix, where $M=NFFT/2+1$. In Matlab, it is constructed in line 62. Intuitively, it compresses high frequencies and differentiates at speech-related frequencies, so it can differentiate speeches more correctly.

$H$ is the mel filterbank matrix, whose rows are filters centered at $M$ different frequency points such that

* Each window $W\_k$ is a shifted triangular function with peak frequency $f\_k$
* $W\_k[f\_{k-1}] = W\_k[f\_{k+1}] = 0$ (i.e. the magnitude of each window vanishes at its neighbor’s peak frequency)
* The first peak frequency $f\_0 = 0$
* The last peak frequency $f\_M = F\_s/2$, where $F\_s$ is the Nyquist frequency
* The peak frequencies $f\_k$ distribute linearly on the mel scale. This means $delta(m) = m\_k - m\_{k-1} = (m\_M - 0) / M = C$, which is a constant, and $m\_k \approx 2593\*\log\_10(1+\frac{f}{700})$