Week 3

Audio Analysis 1

[1] Introduction to Audio Analysis – A Matlab Approach, Theodoros Giannakopoulos and Aggelos Pikrakis
[2] Machine Learning for Audio, Image and Video Analysis, Francesco Camastra and Alessandro Vinciarelli
[3] Introduction to Digital Speech Processing, Rabiner and Schafer

How do you analyze continuous media?

Short-term overlapping windows!

$$x_i(n) = x(n)w(n-m_i)$$

Windowing Issues

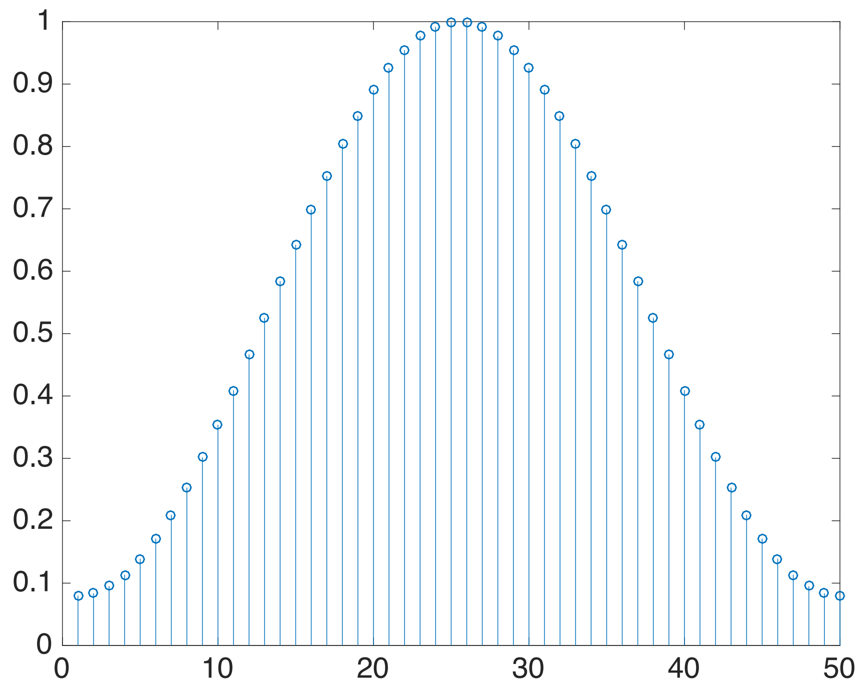
1. What should be the shape of the window?
2. What should be the duration of the window?
3. How much should be overlap between two consecutive windows?

Choosing Window Shape

- Windowing distorts frequency response (spectral leakage)
- With rectangular window, additional high frequency components appear
- Choose a shape that causes least distortion

Choosing Window Shape

- Rectangular
- Hanning
- Hamming
- Blackman
- Kaiser



Choosing Window Size

- Smaller window provides better time resolution
- Bigger window provides better frequency granularity, but loses time resolution
- We generally choose 10ms to 50ms for audio analysis

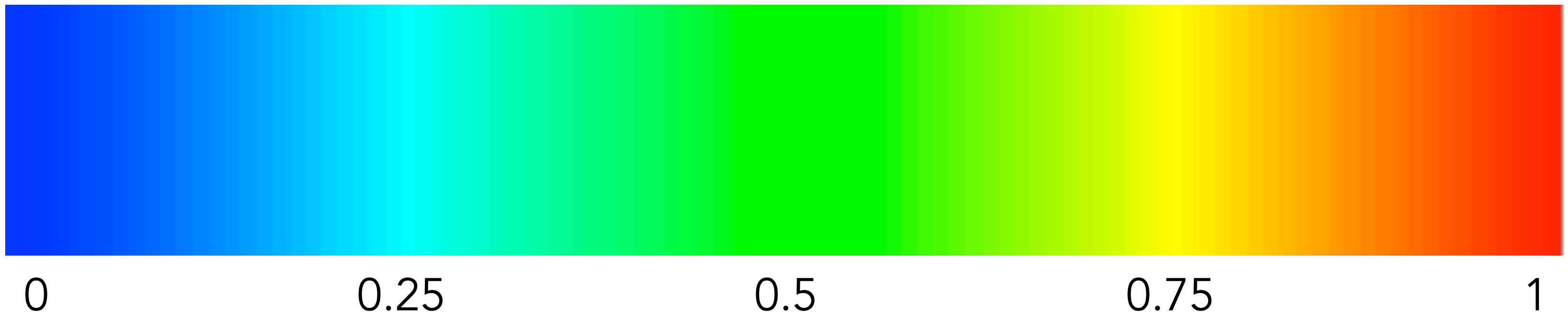
Choosing overlap

- Overlap also improves better time resolution without affecting frequency response, but needs more resources
- Experimentally choose the overlap needed for the given task
- Generally we choose 50% overlap

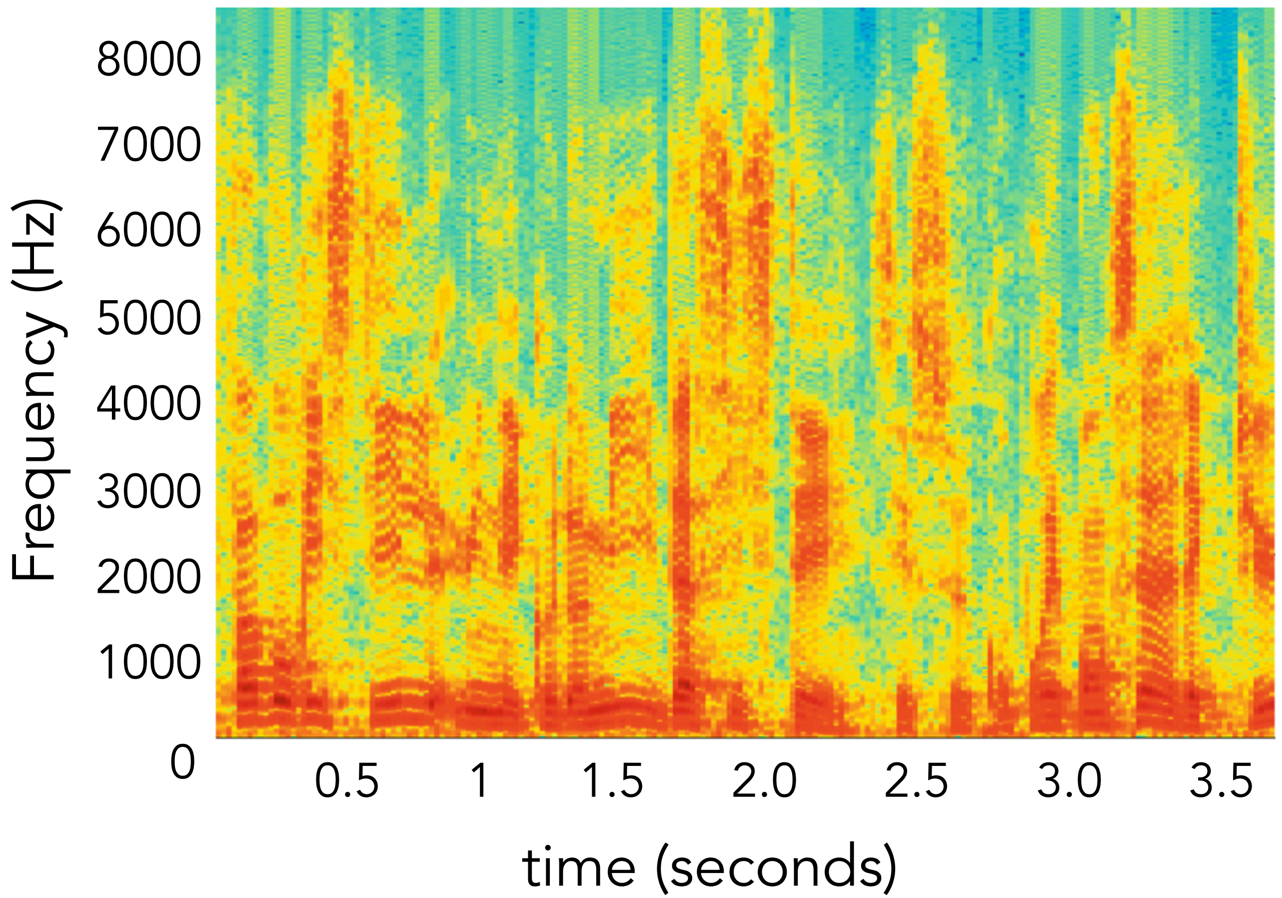
Mid-Term Windowing (1s-10s)

- Audio signal is first divided into mid-term segments
- Shot-term processing is done on each segment
- Effectively, it is like combining few shot-term coefficients

Color Bar Representation of Magnitude



Spectrogram



What are features?

- Abstract representation of the signal
- Features should be distinctive
- Features should be compact

Time Domain Vs Frequency Domain

- Time domain features process the signal directly
- Frequency domain features are derived from the frequency response of the signal

Time Domain Features

How to detect silence in audio?

The sample magnitudes are low during silence!

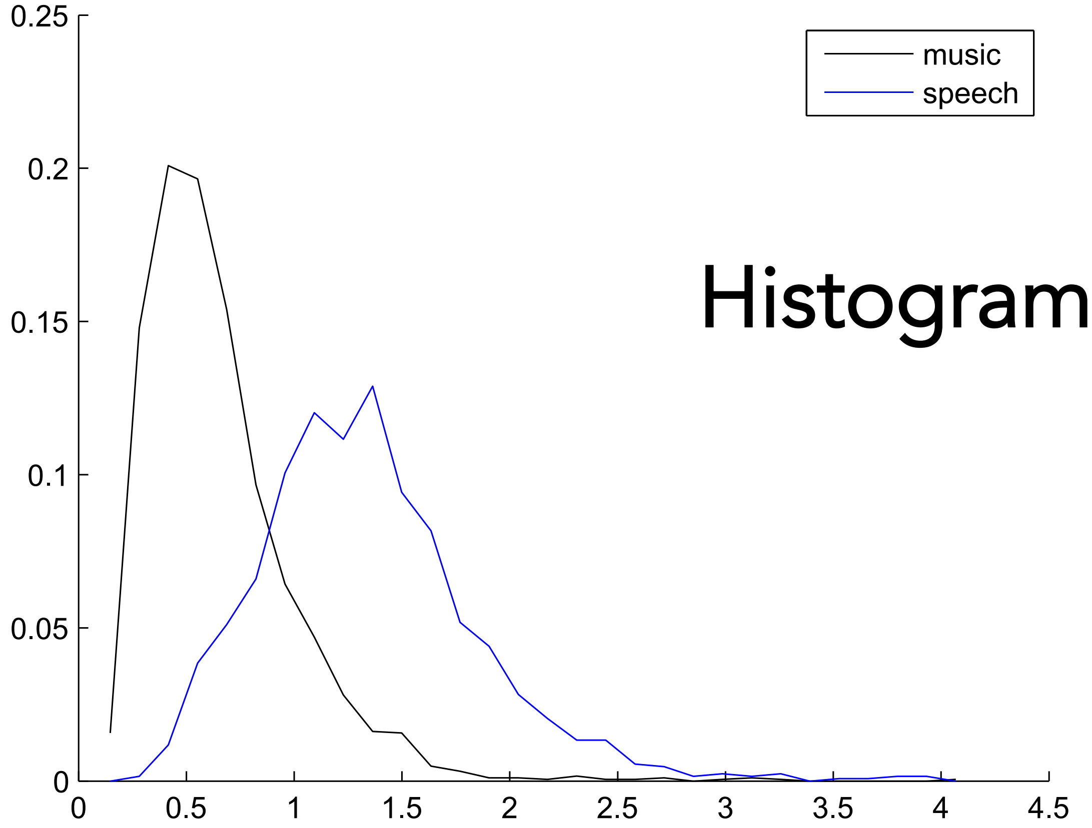
Audio Energy

$$E(i) = \frac{1}{W_L} \sum_{n=1}^{W_L} |x_i(n)|^2$$

Audio Energy Applications

- Silence has low energy
- Speech has more energy variance than music, why?
- How to compare variance of energy of two segments?

We can normalize σ^2 by μ



Histogram of σ^2 / μ

Given two audio files, how will you decide which file is more noisy?

Zero Crossing Rate

The rate of sign changes

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} | \, sgn[x_i(n)] - sgn[x_i(n-1)] \, |$$

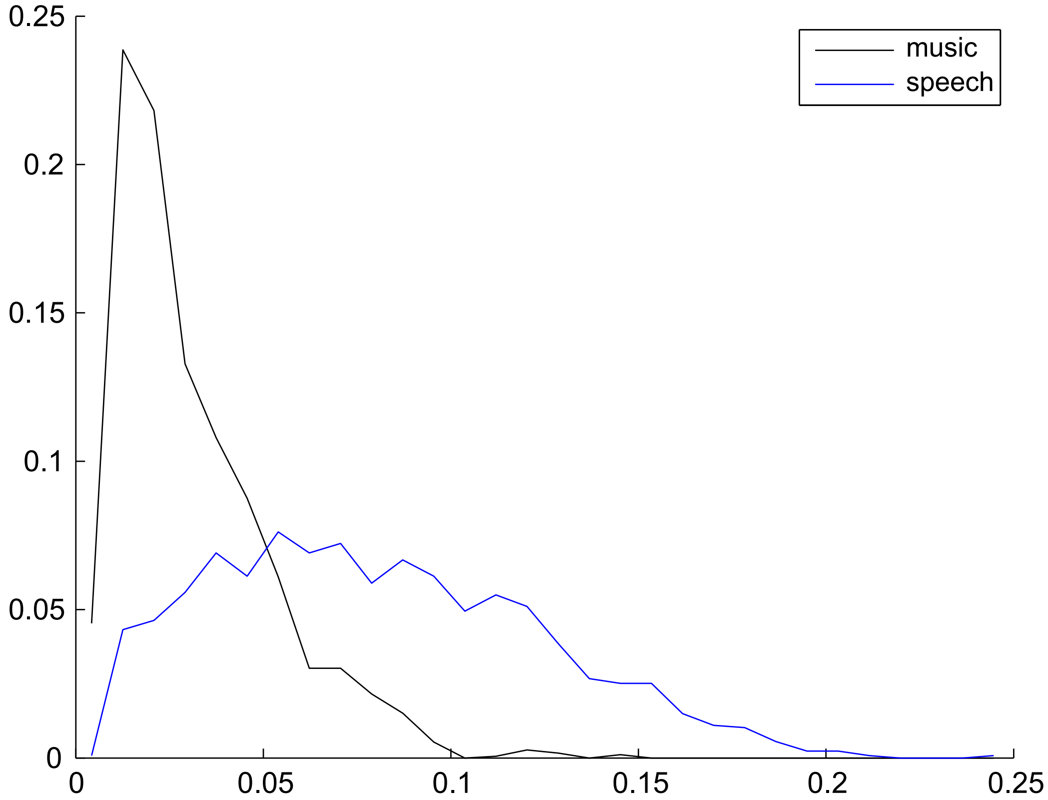
where

$$sgn[x_i(n)] = \begin{cases} 1, & x_i(n) \geq 0, \\ -1, & x_i(n) < 0. \end{cases}$$

Properties of ZCR

- Generally noise/silence/unvoiced speech has higher ZCR than voiced speech
- High ZCR implies high frequency in a coarse manner
- Variance of ZCR is higher for speech than music

Histogram of σ of ZCR



How would you capture smooth and abrupt variations in audio sample?

e.g. gunshot

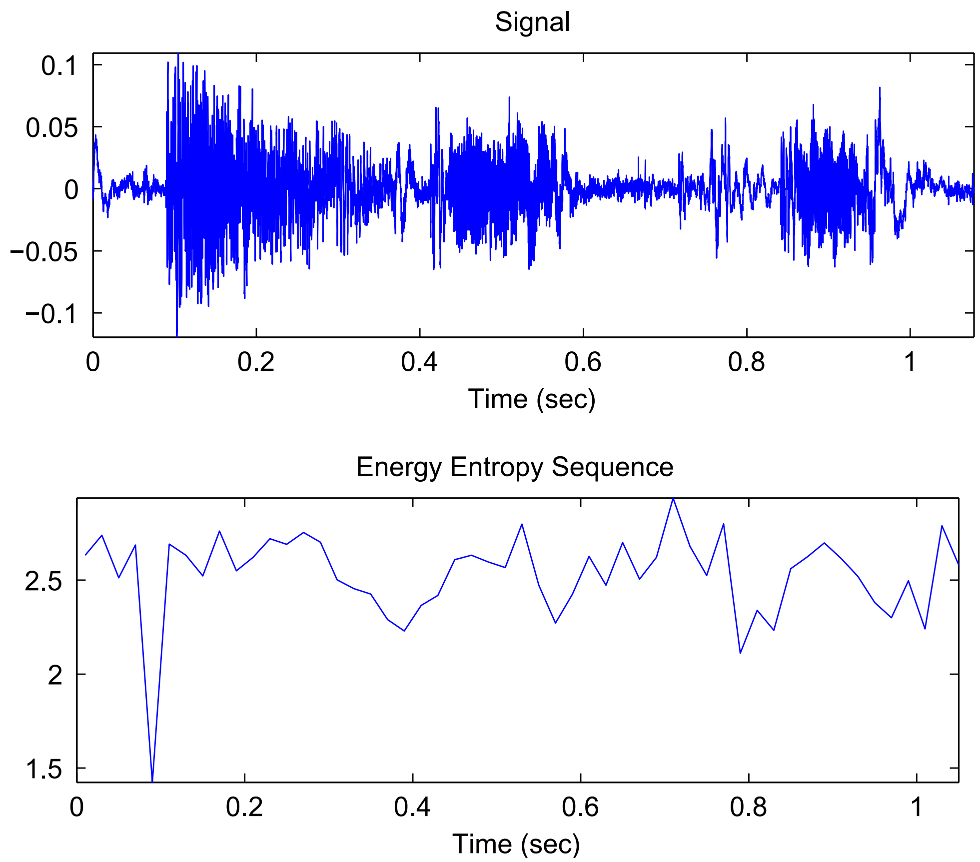
Entropy of Energy

$$e_j = \frac{E_{subFrame_j}}{E_{shortFrame_i}}$$

$$E_{shortFrame_i} = \sum_{k=1}^{K} E_{subFrame_k}$$

$$H(i) = -\sum_{j=1}^{K} e_j \cdot \log_2 (e_j)$$

Entropy reduces at the onset of three gunshots



Properties of energy entropy

- Both short-term and long-term analysis are possible
- Low entropy at onset of many sounds, e.g. gunshot, explosion
- Generally lower values for electronic music and higher for classic music

Time Domain & Frequency Domain

- Features discussed so far are calculated in time domain
- Sometimes frequency components are more informative

MFCC

Mel-Frequency Cepstrum Coefficients

or

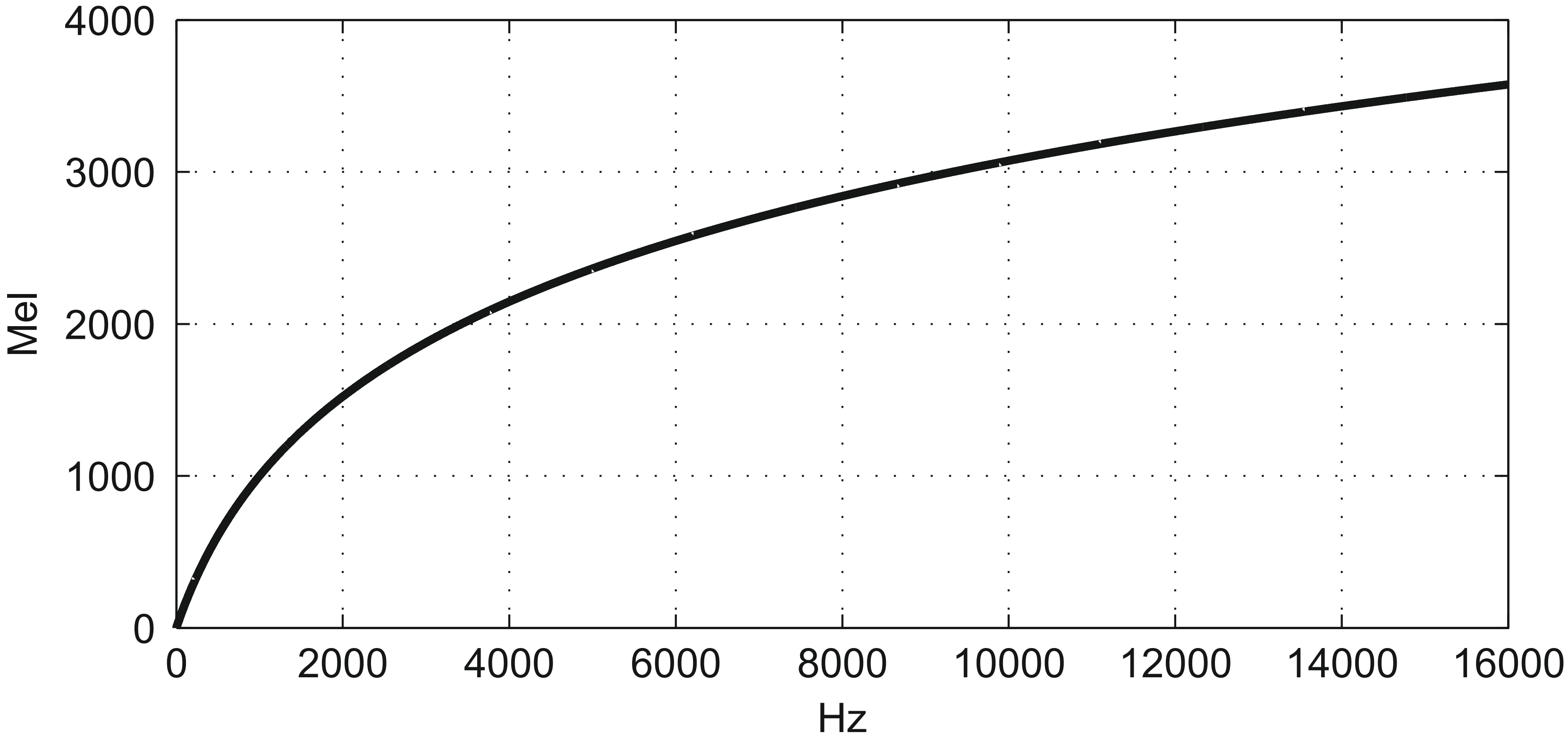
Mel-Frequency Cepstral Coefficients

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

Main Observation

Human Auditory Systems can distinguish neighboring frequencies better in lower region!

Mel-Scale



4000
3000
2000
1000
0
Mel
0
2000
4000
6000
8000
10000
12000
14000
16000
Hz

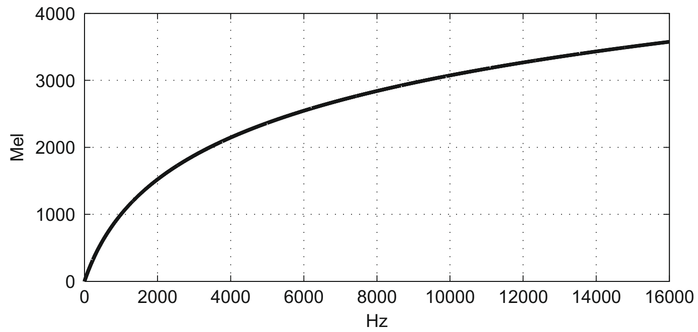
1. Divide the signal into short frames

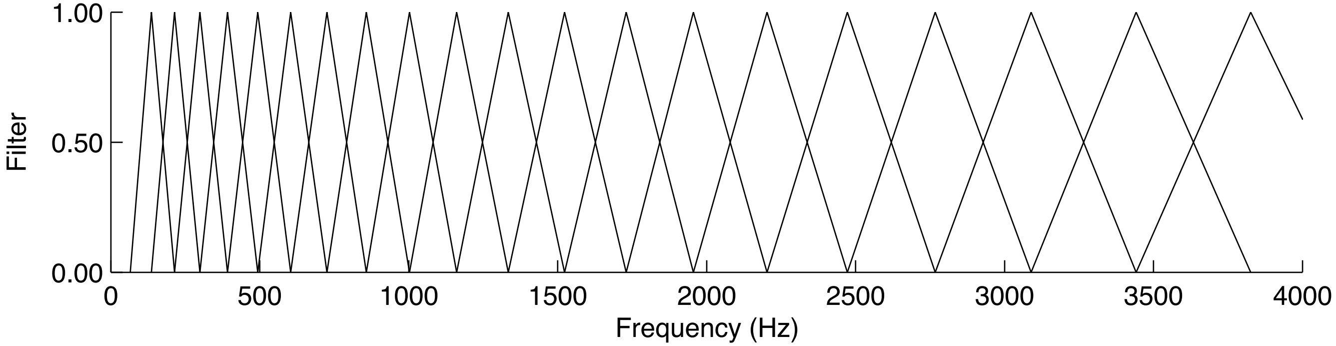
- assume audio does not change statistically in short periods
- generally 20-40ms frames
- frame step is generally 10ms

2. Calculate DFT

- DFT points are more than window size
- for a 400 sample window, take 512 point DFT
- consider only half coefficients, i.e., 257 in the case above
- calculate poser spectral coefficients, which is square of the absolute value divided by total number of coefficients (257)

3. Apply Mel filter-bank (20-26 filters)





4. Determine energy in each filter

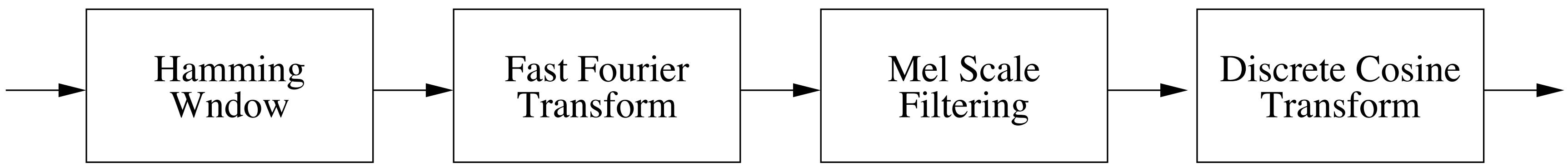
- we have 26 vectors of size 257 each as filter bank
- calculate sum of coefficients in each filter after multiplying with triangular window
- this will lead to 26 values which represent energy in each filter bank

5. Take logarithmic of each filter-bank energy

- we don’t hear loudness on linear scale but on a logarithmic scale

6. Take DCT of log filter-bank energies and keep 2-13 coefficients!

MFCC



Hamming Wndow
Fast Fourier Transform
Mel Scale Filtering
Discrete Cosine Transform

More Spectral Features

- Spectral centroid
- Spectral entropy
- Spectral flux
- Spectral rolloff