MFCC tutorial

Brought to you by EE6641 TAs

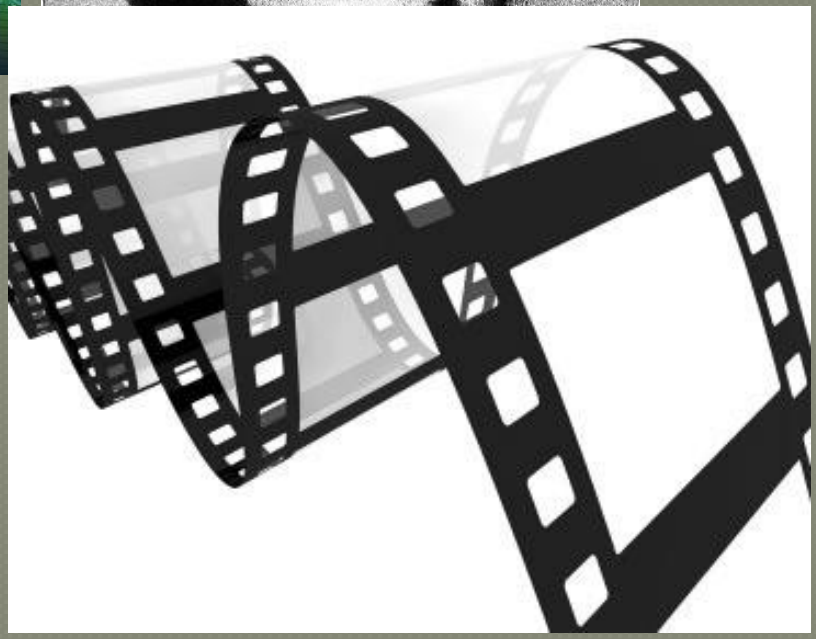
Outline

- History
- Mel frequency
- Cepstrum
- MFCC
- Applications
- Conclusions

Outline

- History
- Mel frequency
- Cepstrum
- MFCC
- Applications
- Conclusions

Signal Processing



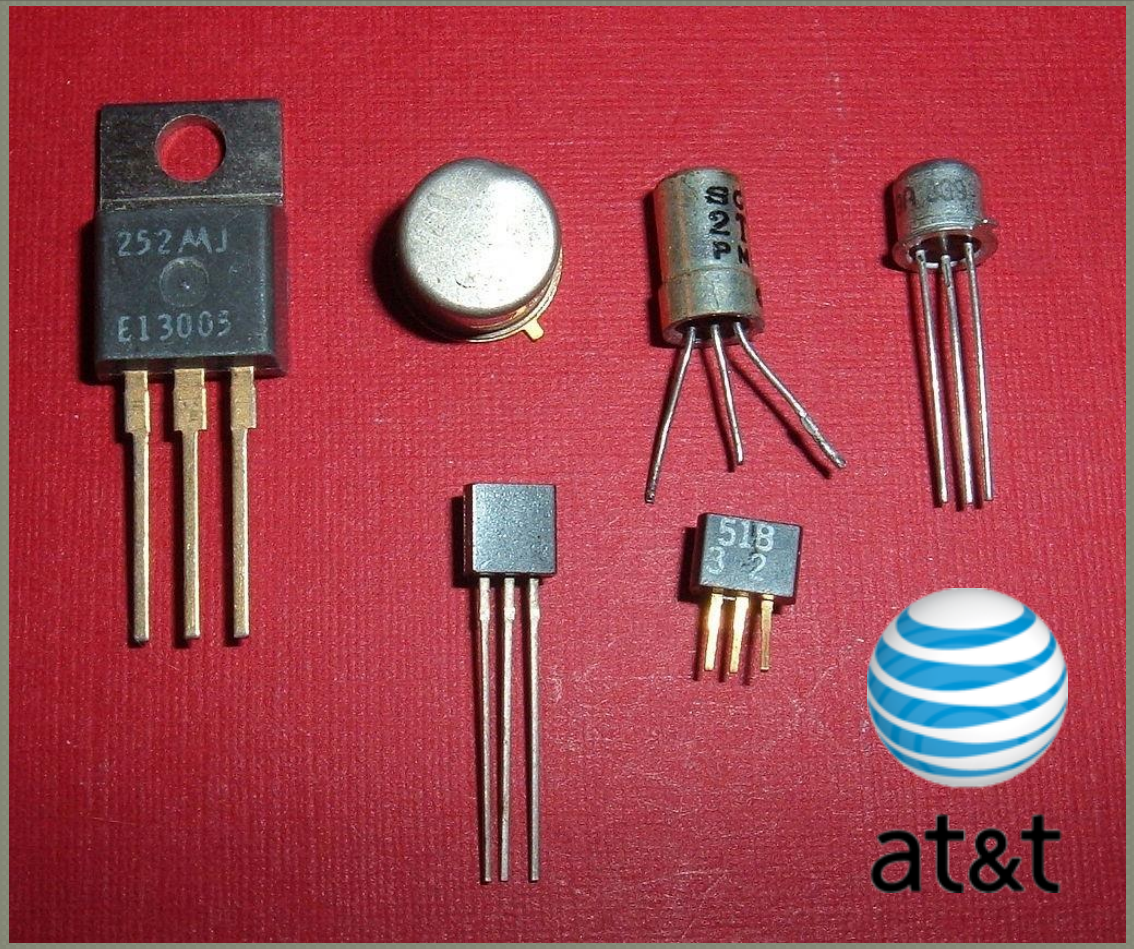
Speech recognition

Pitch detection

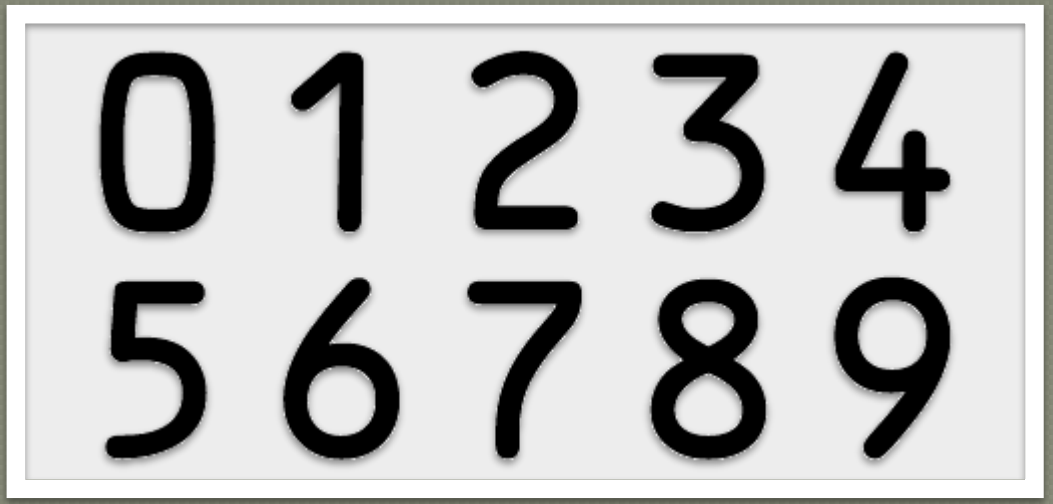
Cover-song detector

and so on…

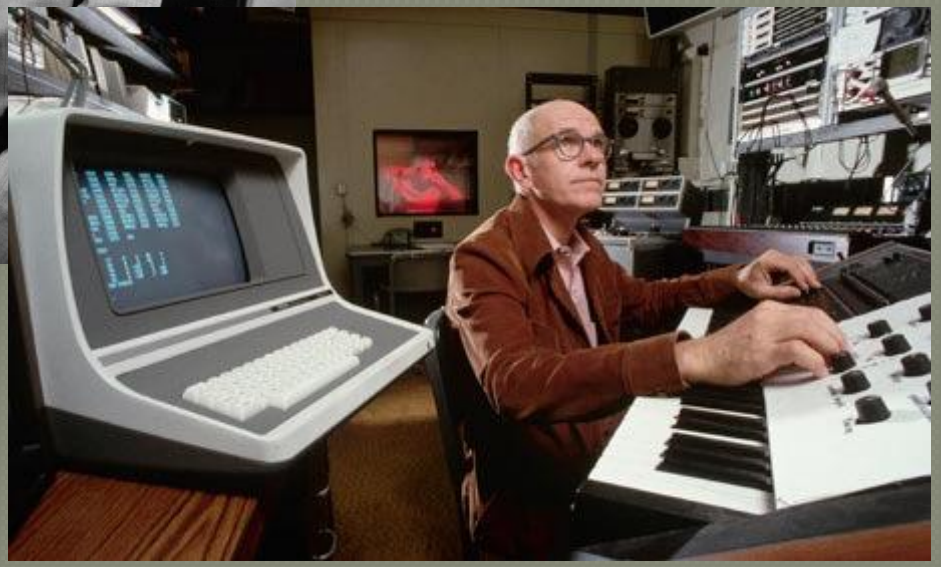
1930s



1950s



1960s

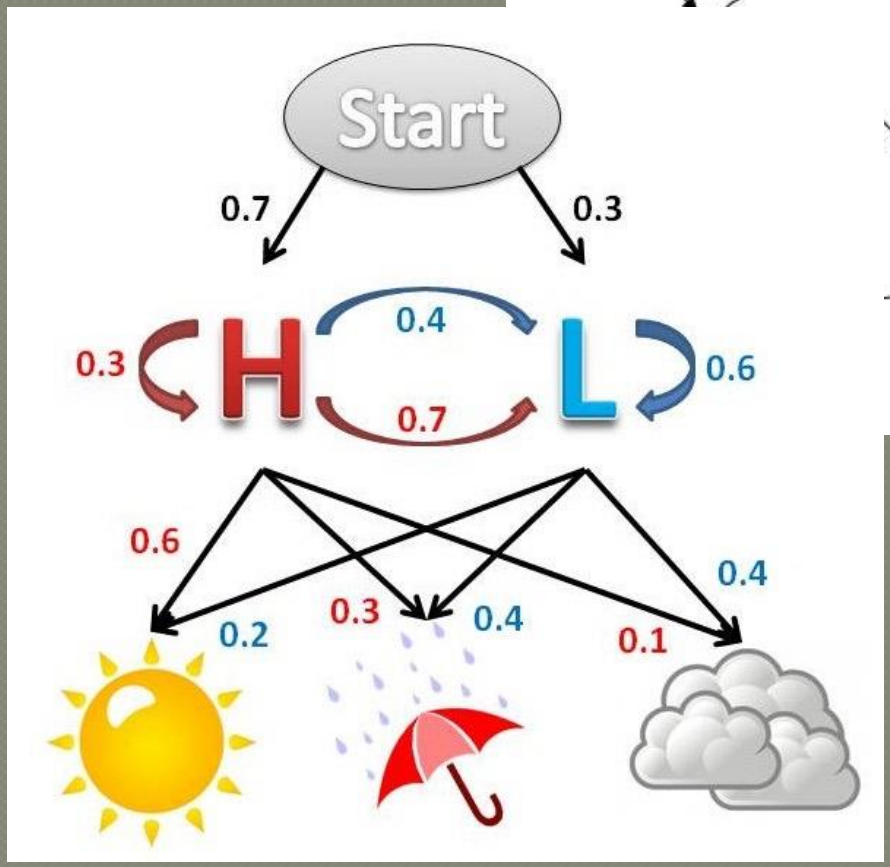


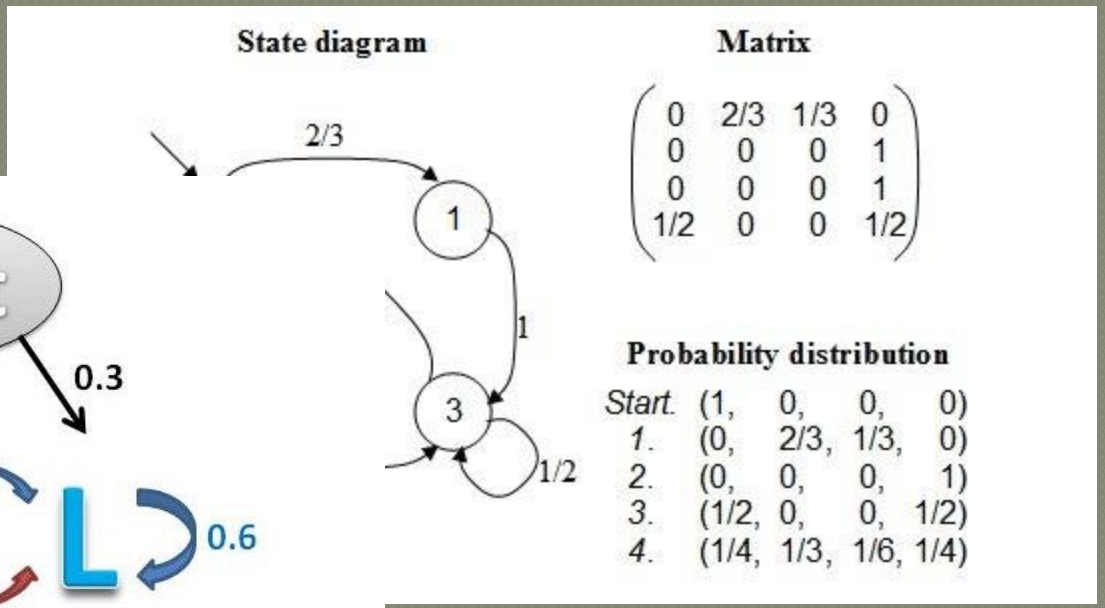
1960s

- Rej Reddy
- Soviet=>DTW capable of 200 words

1960's

- IDA: Leonard





1980s

- Can recognize 2
- 4MB ram => 30s minutes

1990

- Commercial opportunities
- Number of words bigger than human's

2000s

- Lern & Hauspie
- Dragon System
- Later Bankrupt

2010s

- Deep learning
- Reduced 30% error
- “the most dramatic change”

Outline

- History
- Mel frequency
- Cepstrum
- MFCC
- Applications
- Conclusions

Mel-frequency

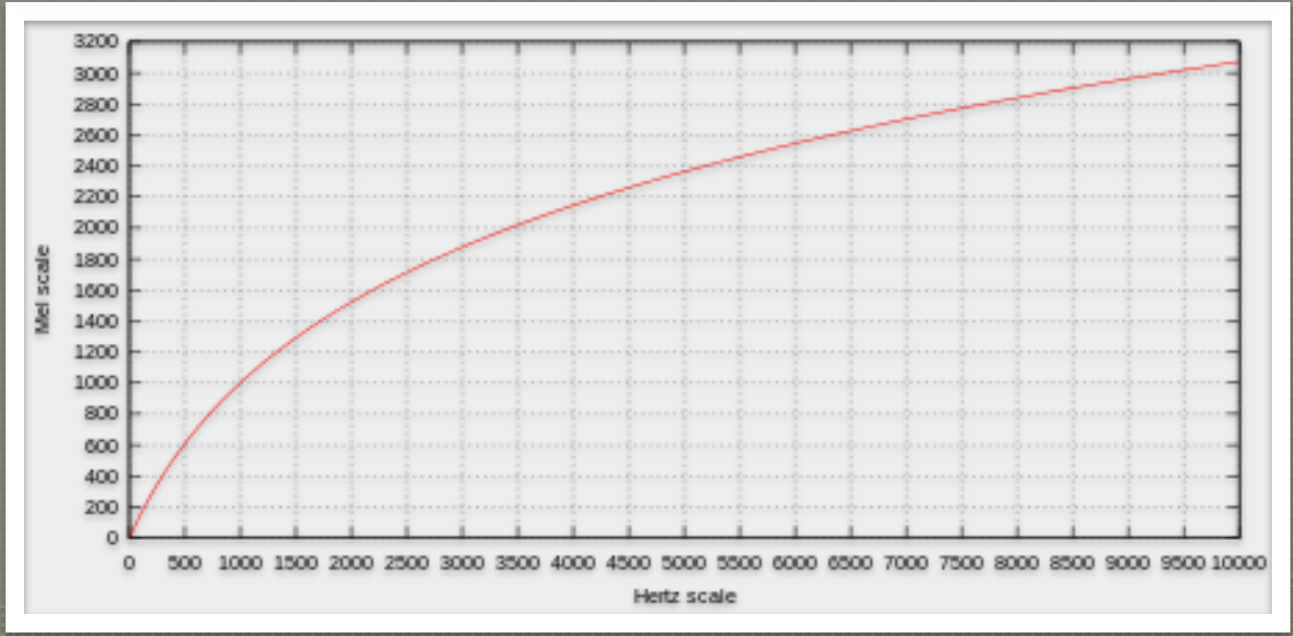
- perceptual scale of pitch
- 1000 to 1000
- "聽閾"
- Not all equations are the same

Mel-frequency

Hz	40	161	200	404	693	867	1000	2022	3000	3393	4109	5526	6500	7743	12000
mel	43	257	300	514	771	928	1000	1542	2000	2142	2314	2600	2771	2914	3228

$$m = 2595 \log_{10}(1 + \frac{f}{700}) = 1127 \log_e(1 + \frac{f}{700})$$

$$f = 700\left(10^{\frac{m}{2595}} - 1\right) = 700(e^{\frac{m}{1127}} - 1)$$

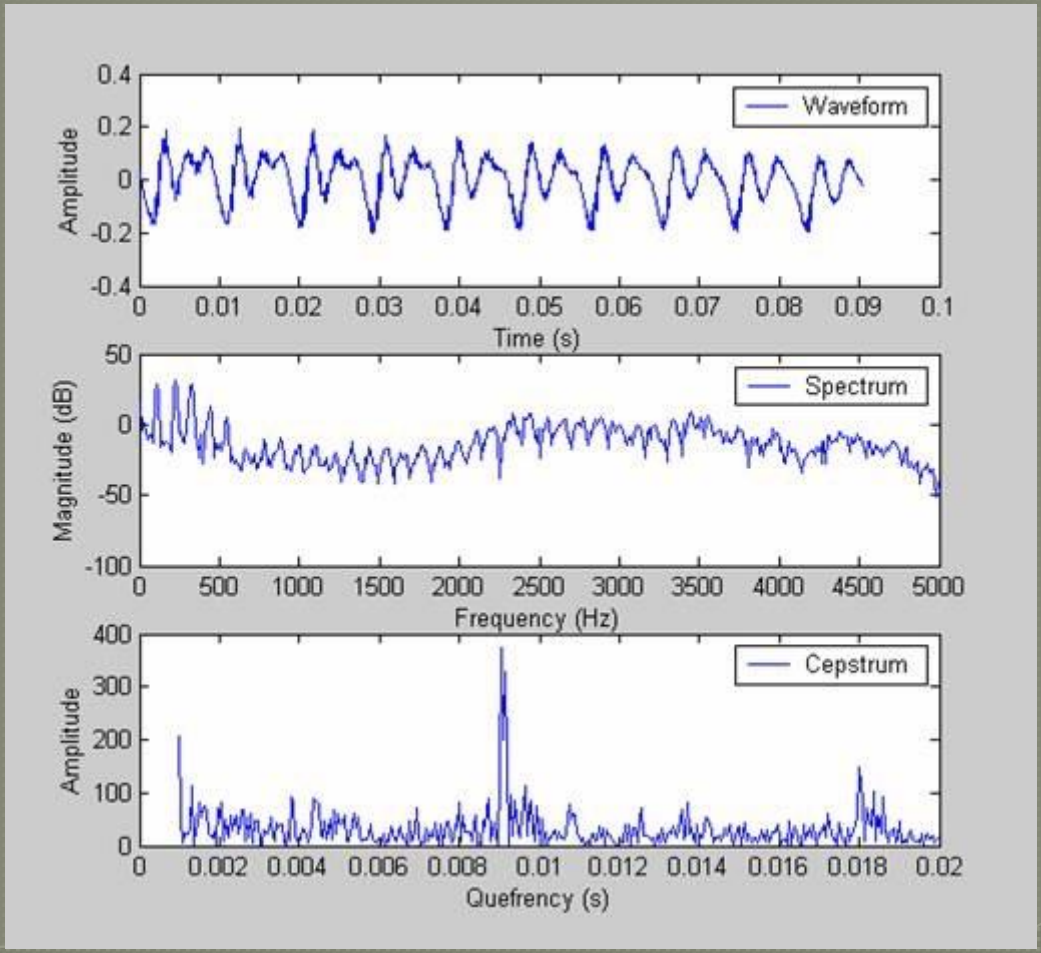


Outline

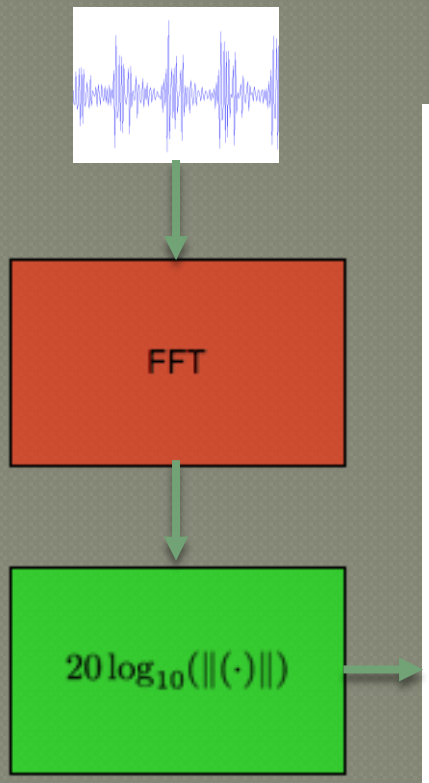
- History
- Mel frequency
- Cepstrum
- MFCC
- Applications
- Conclusions

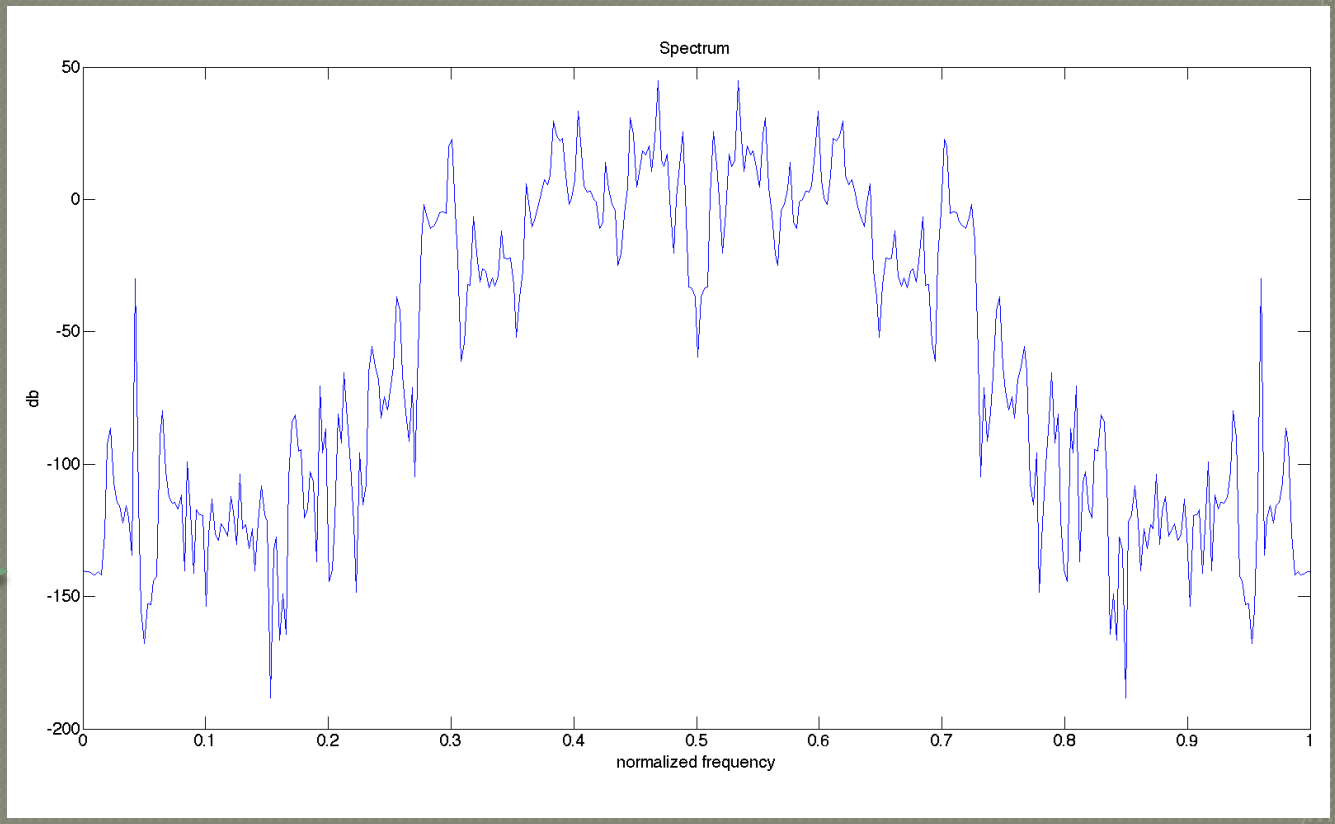
Cepstrum

- FFT => abs() => log() => IFFT(FFT)
- “quefrency”

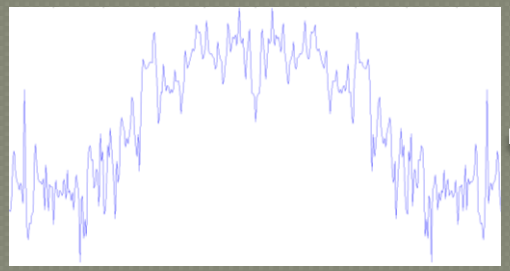


Spectral Envelope

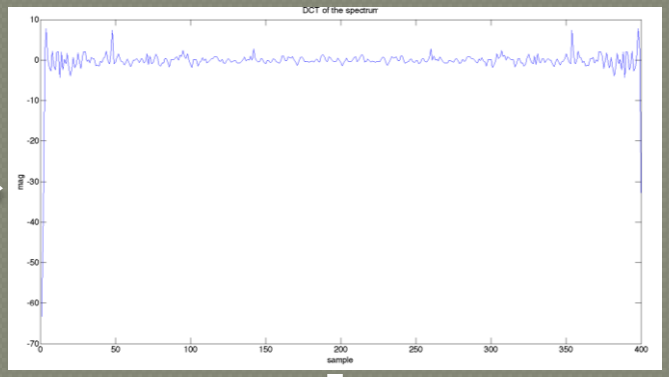




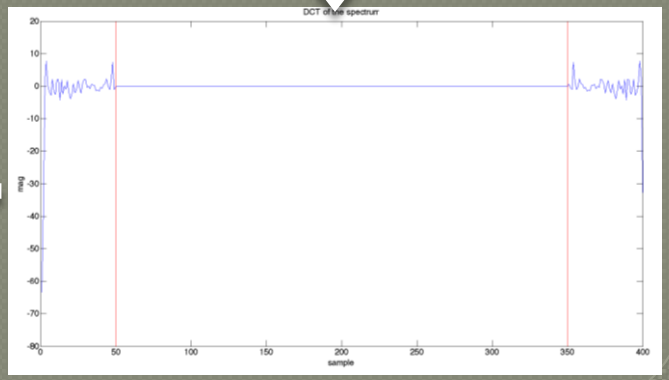
Spectral Envelope



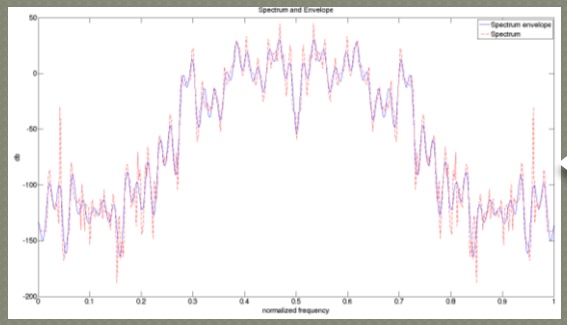
FFT



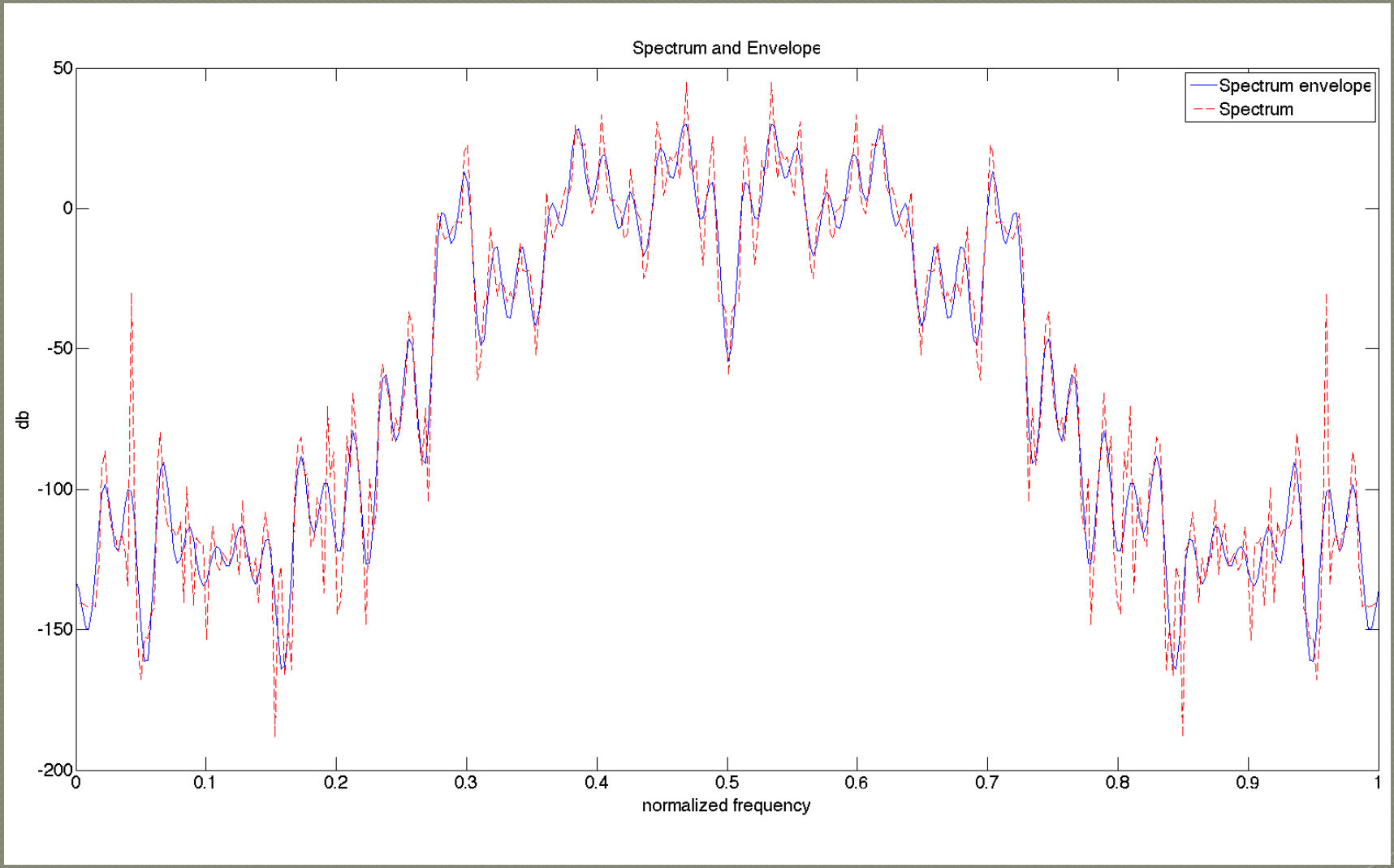
Low-pass



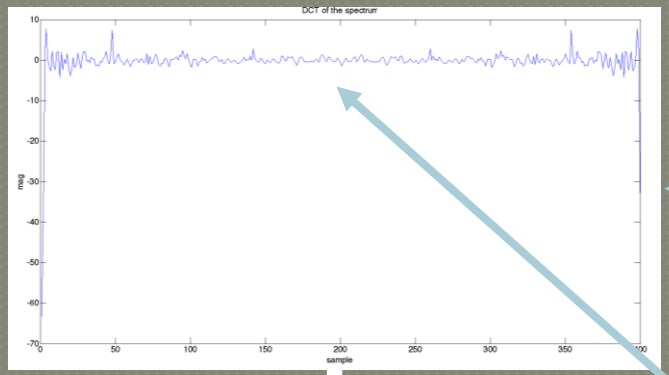
IFFT



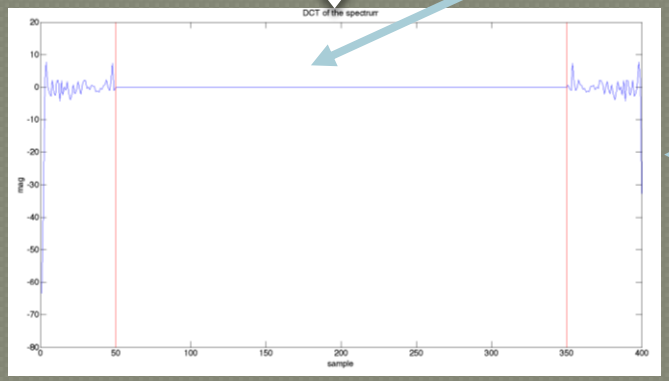
Spectral Envelope



Why dB?



Low-pass



$$\log(\|X\|)$$

$$=$$

$$\log(\|E\|)$$

$$+$$

$$\log(\|H\|)$$

$$X = E \cdot H$$

$$\Downarrow$$

$$x = e * h$$

Disc-Cosine-Trans v.s. Disc-Fourier-Trans

DCT

Difference?

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

DFT

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

Disc-Cosine-Trans v.s. Disc-Fourier-Trans

DCT

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

2x resolution

0.5x memory

O(N*N)

DFT

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

Disc-Cosine-Trans v.s. Disc-Fourier-Trans

$$X_C[k] = \sum_{n=0}^{N-1} x[n] \cos(\frac{\pi}{N} nk)$$

$$= \sum_{n=0}^{N-1} x[n] \cos(\frac{2\pi}{2N} nk)$$

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

$$= \frac{1}{2} \sum_{n=0}^{N-1} (y[n] + y[2N-1-n]) \cos(\frac{2\pi}{2N} nk) \quad y[n] = [x \ \mathrm{rev}(x)]$$

$$= \frac{1}{2} \sum_{n=0}^{2N-1} y[n] \cos(\frac{2\pi}{2N} nk)$$

$$= \frac{1}{2} \mathrm{Re}\{Y_F[k]\} = \frac{1}{2} Y_F[k] \quad k \in [0, \ N-1]$$

Disc-Cosine-Trans v.s. Disc-Fourier-Trans

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

$$= \mathrm{Re}\left\{ \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{2N}kn} \cdot e^{-j\frac{\pi}{2N}k} \right\}$$

$$= \mathrm{Re}\left\{ \sum_{n=0}^{2N-1} y[n] e^{-j\frac{2\pi}{2N}kn} \cdot e^{-j\frac{\pi}{2N}k} \right\} \quad y = [x\ 0]$$

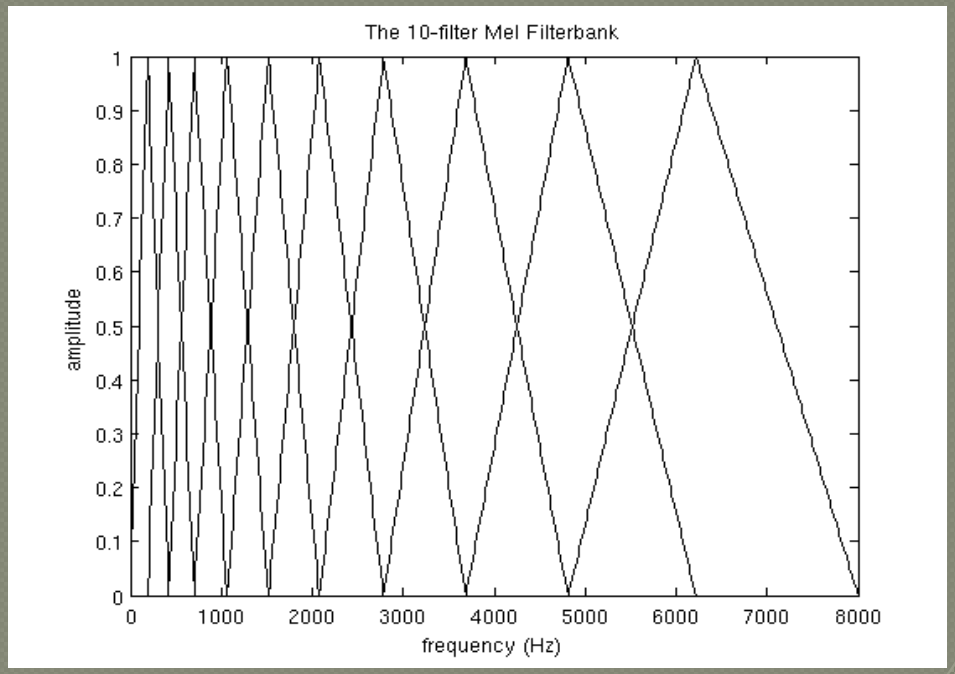
$$= \mathrm{Re}\left\{ Y_F[k] \cdot e^{-j\frac{\pi}{2N}k} \right\} \quad k \in [0,\ N-1]$$

Outline

- History
- Mel frequency
- Cepstrum
- MFCC
- Applications
- Conclusions

MFCC

- FFT => power spectrum =>
- triangular filter banks (usually 26)
- log => DCT(IDCT)
- 取係數 (usually 13)



MFCC

- Why MFCC?
- Simplicity (Only several coefficients)
- Smoothness

Outline

- History
- Mel frequency
- Cepstrum
- MFCC
- Applications
- Conclusions

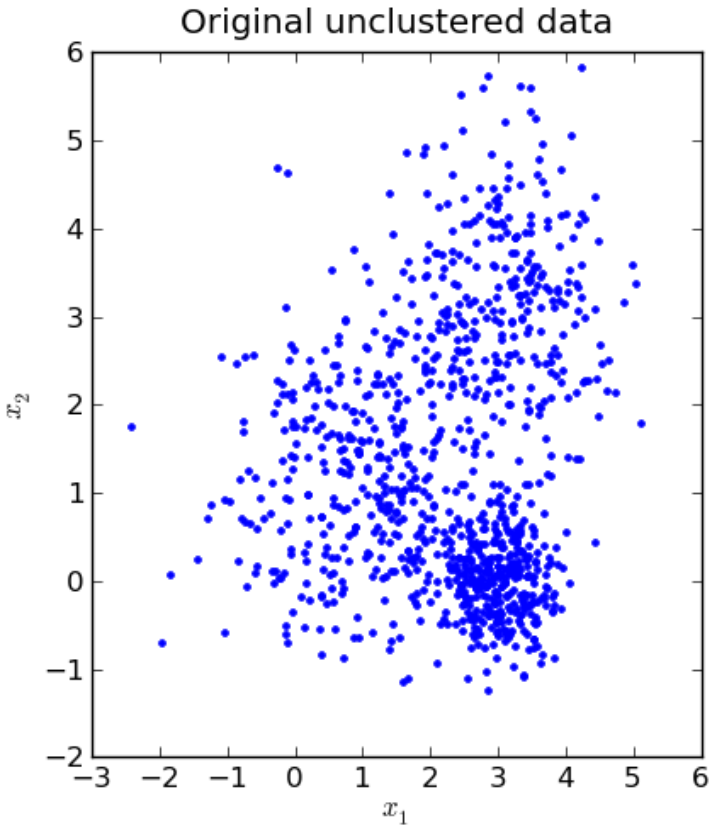
Machine Learning

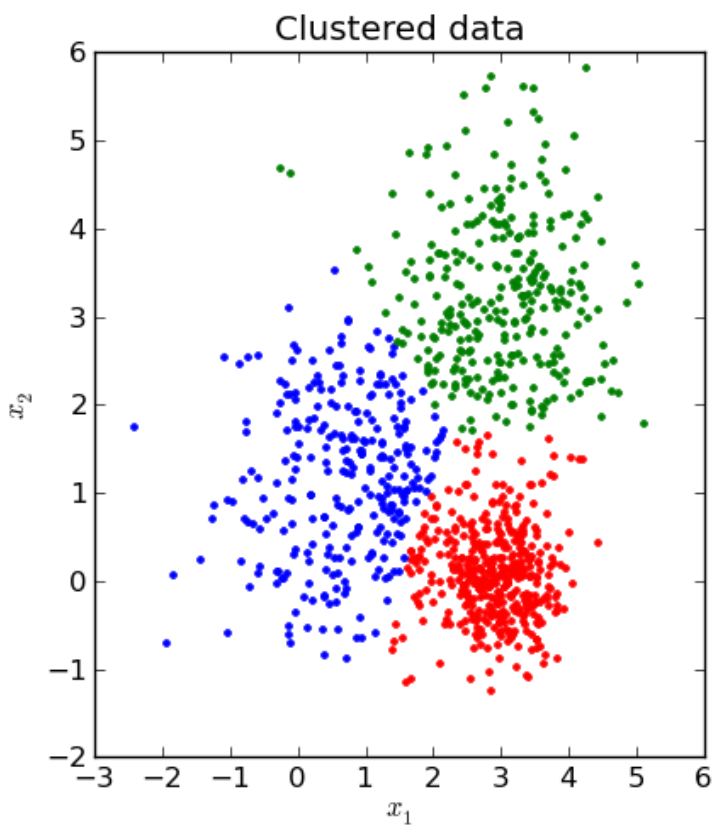
- Unsupervised learning
- Supervised learning
- Semi-supervised learning

Unsupervised learning

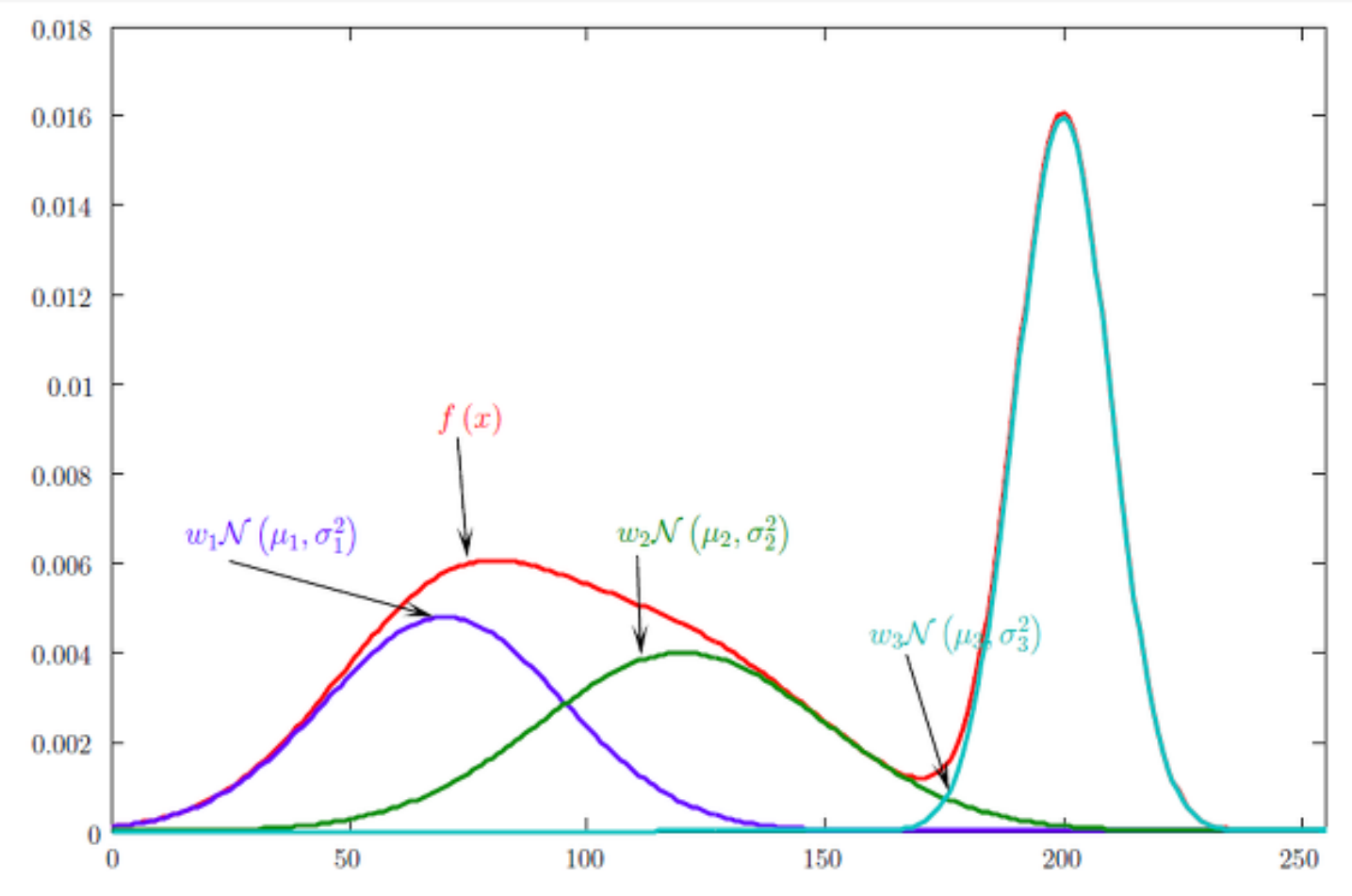
- Expectation maximization
- E-step vs M-step

K-means





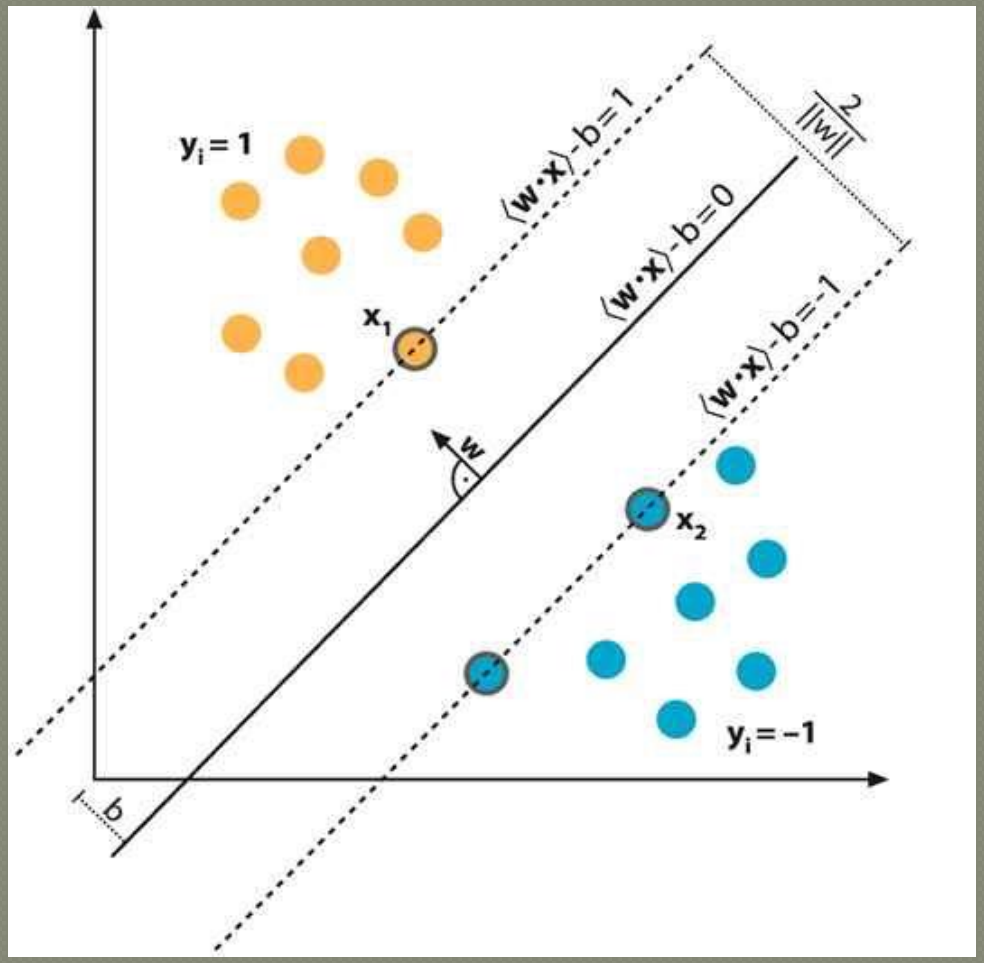
GMM


0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
50
100
150
200
250
$f(x)$
$w_1\mathcal{N}(\mu_1, \sigma_1^2)$
$w_2\mathcal{N}(\mu_2, \sigma_2^2)$
$w_3\mathcal{N}(\mu_3, \sigma_3^2)$

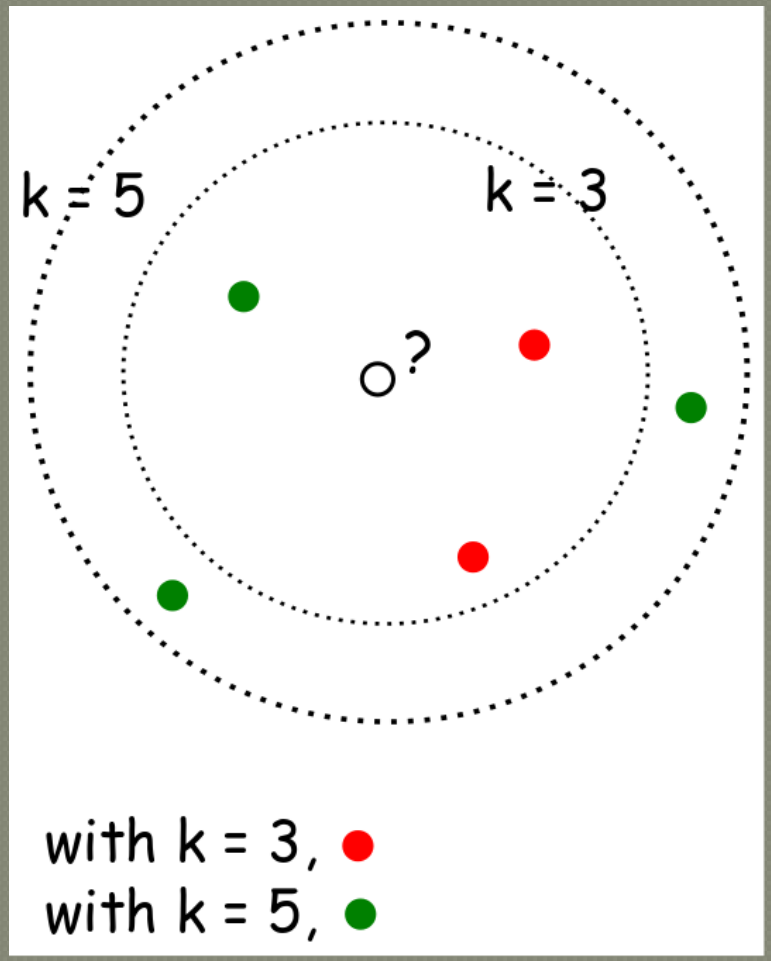
Supervised learning

- Based on “labels”
- Empirical vs General
- Error minimization

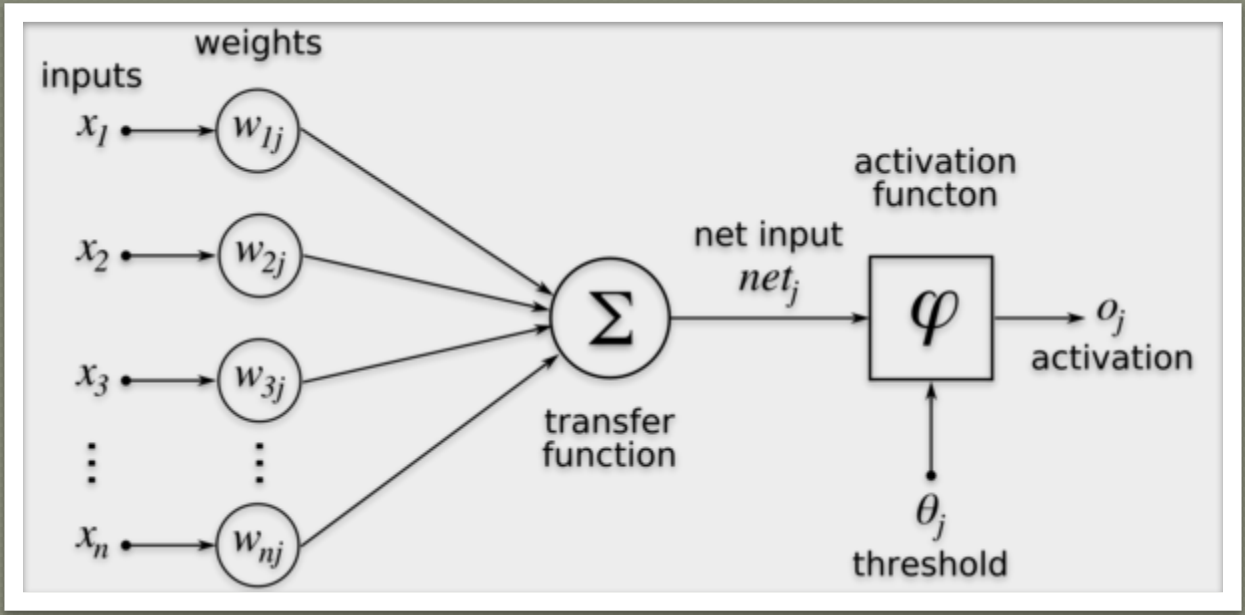
SVM


$y_i = 1$
$\langle w \cdot x \rangle - b = 1$
$\frac{2}{\|w\|}$
$\langle w \cdot x \rangle - b = 0$
$\langle w \cdot x \rangle - b = -1$
x_1
x_2
w
b
$y_i = -1$

kNN



Neural nets


weights
inputs
x_1
x_2
x_3
x_n
w_{1j}
w_{2j}
w_{3j}
w_{nj}
Σ
transfer function
net input
net_j
activation functon
φ
o_j
activation
θ_j
threshold

Musical Instruments Identification

- Use audio recorded by ourselves





Outline

- History
- Mel frequency
- Cepstrum
- MFCC
- Applications
- Conclusions

Quick Recap

- **Why Mel-Filterbank?**
  - 人耳聽覺
- **Why DCT?**
  - 頻譜對稱性
  - 兩倍的解析度
- **Why dB?**
  - 系統分解

Conclusions

- **What's next?**
- 工欲善其事，必先利其器。

Any Questions?