Speech Tech

Week-2

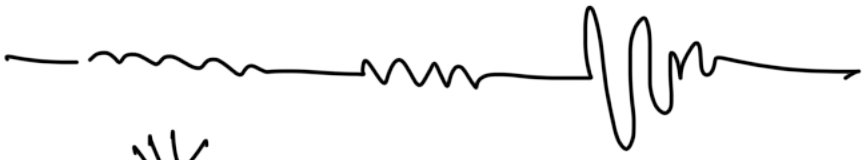
L1 Speech Product", Percep" & Freq. Analysis

- speech produc" model
  - ↳ Terminal analogous model
  - ↳ it is similar to the LTI systems, it interacts with input.)
    - linear time invariant filter.

- STFT (using the window of small size)

  in small window, the fequency is quasi-stationary
  - ↳ this is an assump".

⇒ this window moves with some overlap.



we get the fourier transform. → it will be DFT.

- widespread spectrogram ← → narrow spectrogram

→ small window → large

→ good time resoln → poor

→ resonant freq. envelopes the spectrum → pitch harmonics

- In speech, window len = 25 ms
  overlap = 10 ms

- Speech percepn → actual & perceived speech attributes, mel scale, masking

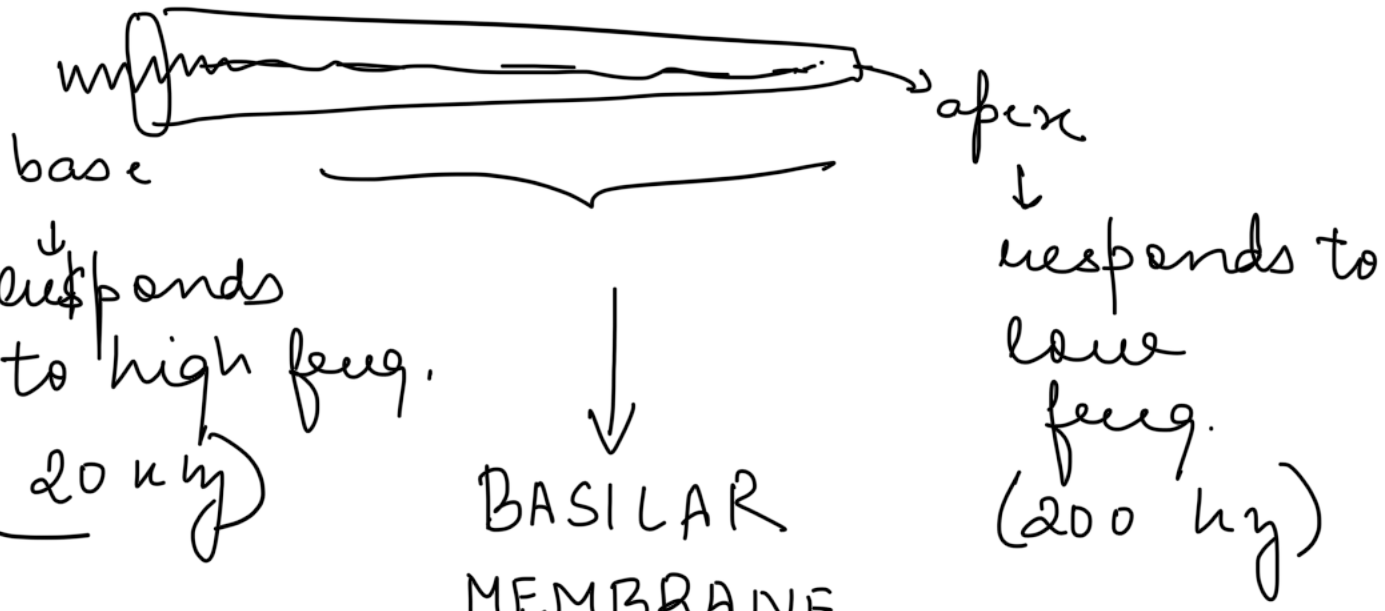
-

Physical attribute of acoustic signal	auditory sys. →	Psychophysical observation
→ intensity		→ loudness
→ freq.		→ pitch.

- pitch & freq.
  → perceptual counterpart of freq.
  → units is mel
  → pitch of a 1000 hz is taken as 1000 mels.

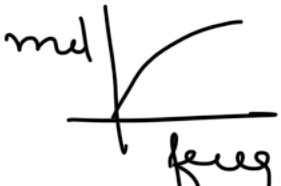
→ obtained by the freq. as perceived by the listener as the signal freq. keep varying.

- the basilar membrane free over body acts as the frequency analyzer



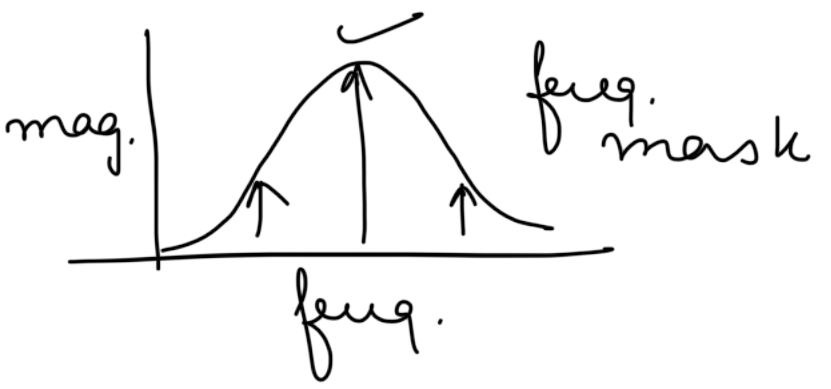
L 2 Perceptual Masking, Cepstrum & Filtering

- MEL scale – quasi-log func^{n} of acoustic frequency s.t. perceptually similar pitch intervals appear equal in width over the full hearing range.

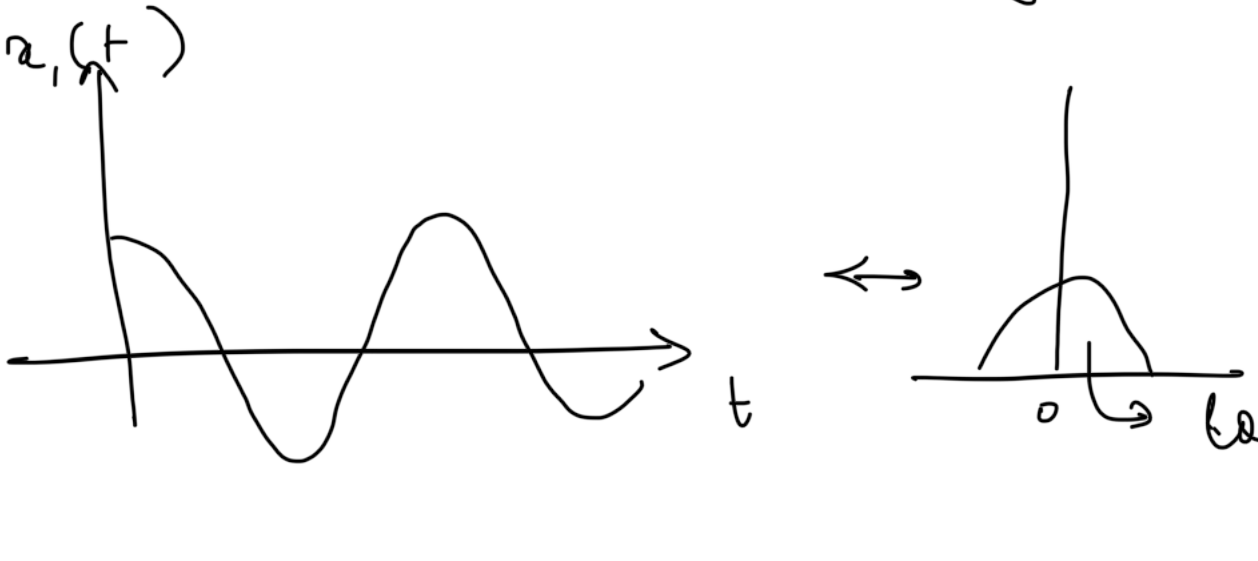


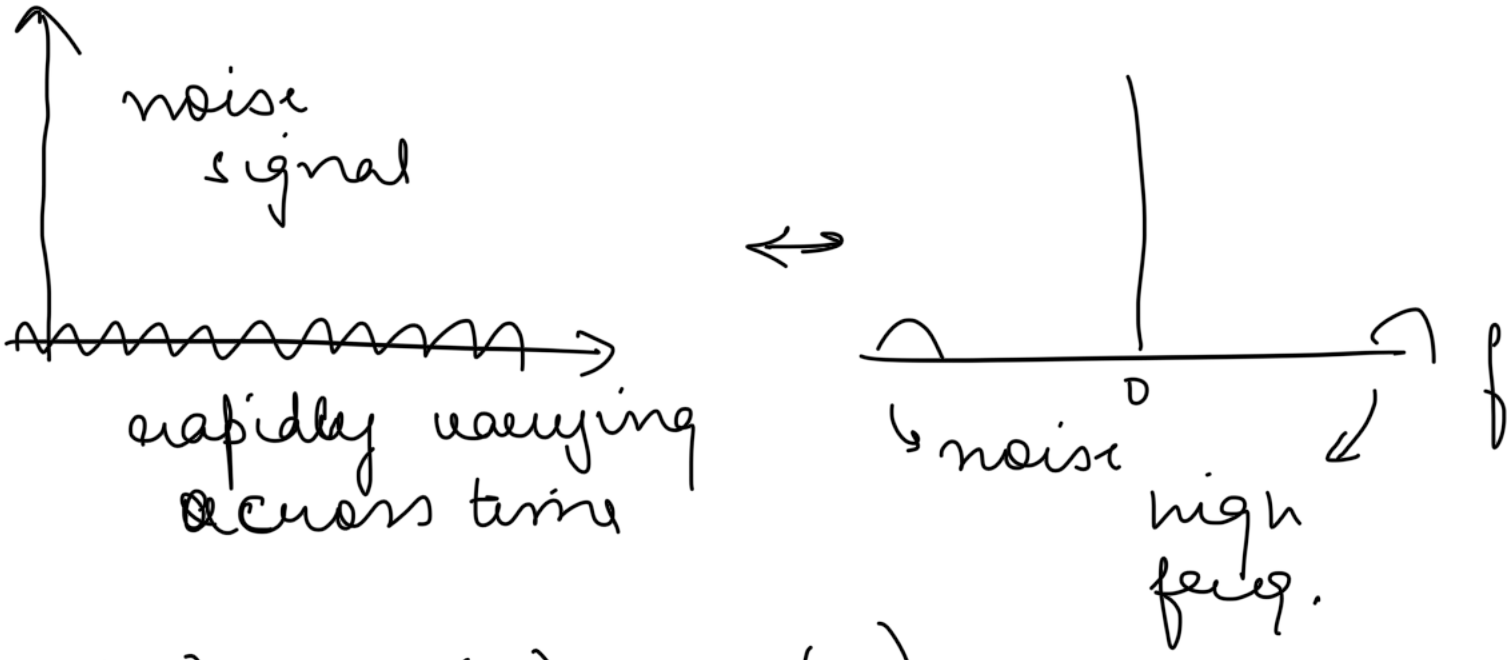
$$Mel = 1127 \log_e \left(1 + \frac{F}{700}\right)$$

- masking of sounds in ∨ of other sounds.

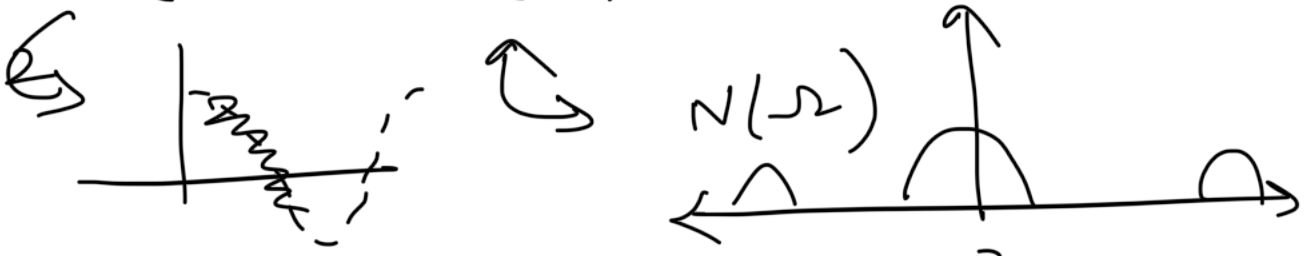


- cepstrum
  - → recall speech produc" model
  - → pitch are the vibrations
  - → speech comes as a result of convolu" of vocal tract response with glottis pulses / turbulent air.
- We have to sep. out the vocal tract response –
  - → yes, can be done.
  - → no, only filtering doesn't work
  - → if 2 signals are added, then filtering works.
  - → what abt. convolved signals?





$$z(t) = x(t) + n(t)$$



from a given $z(t)$, how to get back $x(t)$?

Practical implementation

$$X(\Omega) = Z(\Omega) \cdot H(\Omega)$$

↳ low pass filter

$$\hat{x}(t) = \frac{1}{2T_0} \int_{-T_0}^{T_0} z(t-\tau)\, d\tau$$

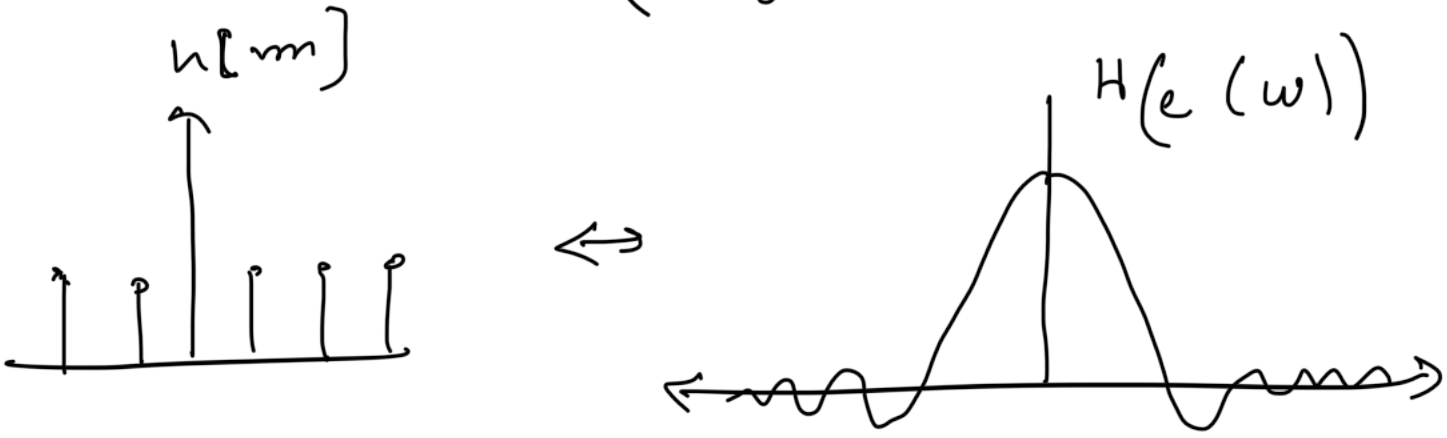
↳ avg. method

The received $\hat{x}(t) \approx x(t)$.

$$\hat{x}[n] = \frac{1}{2M+1} \sum_{m=-M}^{M} x[n+m]$$

$$= \sum h[m]\, x[n-m]$$

$$h[m] = \begin{cases} \frac{1}{2M+1} & -M \le m \le M \\ 0 & \text{otherwise} \end{cases}$$



Cepstrum Analysis

fourier transform of speech signal

we are recovering $x(n)$ from here.

$$\log|S(\Omega)| = \log|V(\Omega)| + \log|E(\Omega)|$$

↳ take out $\log|V(\Omega)|$

yes, by avg. or low-pass filter (LIFT-ering)

L3 Feature Extrac^n for Speech Processing - I

$$|S(\Omega)| = V(\Omega)\ E(\Omega)$$

analogous

$$y[n] = h[n] * x[n]$$

↳ low pass filter

- Convolu^n is multiplica^n in freq. dom.
- Applying log in freq. domain makes the multiplican as addi^n.
- slowly varying vocal-response can then be seperated from a rapidly varying pitch harmonics (instead of filtering, it becomes Liftering)
- now low pass filtering can extract the spectrum related vocal tract response alone. We use a rectangular moving avg. filter.
- features extrac^n → filter bank, MFCC
- to obtain MFCC - a triangular moving avg. filters are used on DFT. They are of not same size & width. Inspired by mel scale

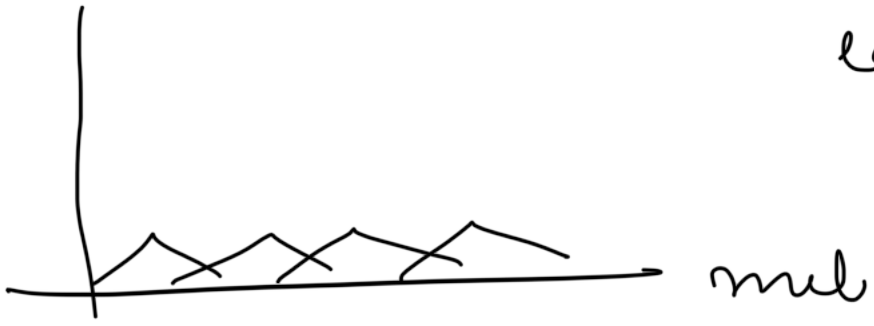
Speech → DFT → Take log → low pass liftering.

- Mel Filter bank - features
- A dCT-2 is performed to decorrelate the features to obtain MFCC.
- basis functions are type of cosine.

$$\Delta MFCC_m[n] = MFCC_m[M] - MFCC_{m-1}[n]$$

L4 Mel Filter

Speech
↓
Pre-Emphasis
↓
Windowing
↓
FFT
↓
Mel Filter Bank → Log
↑
DCT (MFCC)
↑
MFCC + Delta-t Delta Delta



equal spaced
equal B.W.

↓

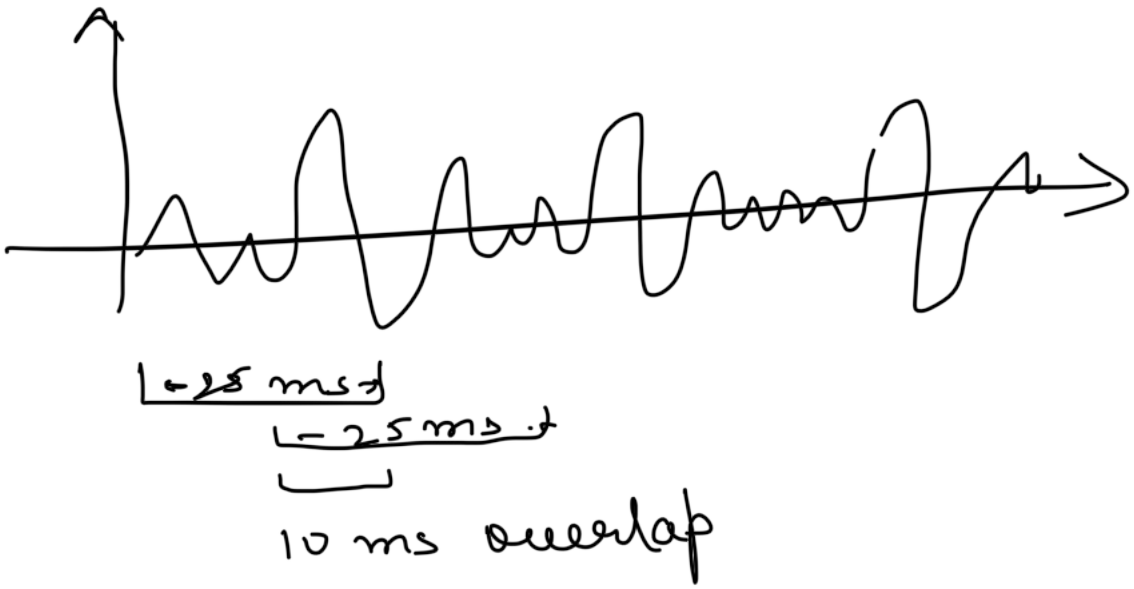
$$f_{mel} = 1125 \ln\left(1 + \frac{f}{700}\right)$$

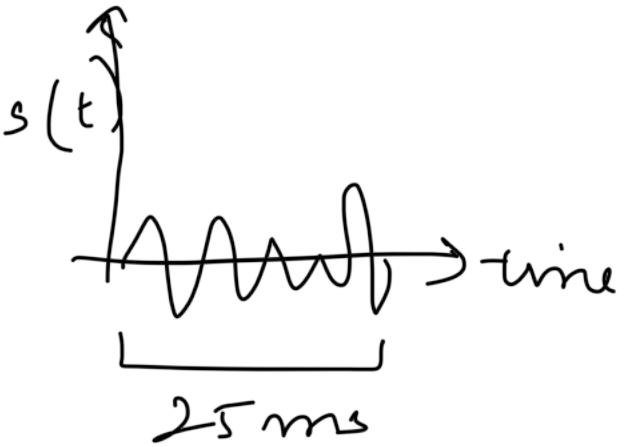
log ⇒ in the magnitude direct log mel filter bank coeff.

40 mel filter ⇒ telephone (8 kHz)

80 mel filter ⇒ wide band speech (16 kHz)

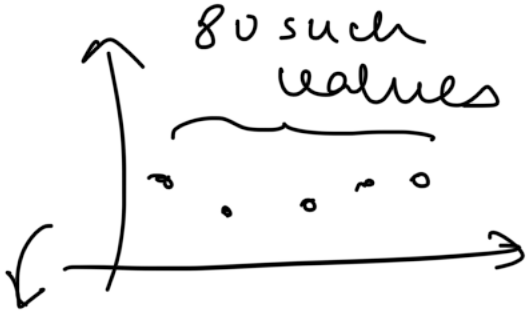
↳ in DL generally this is used.



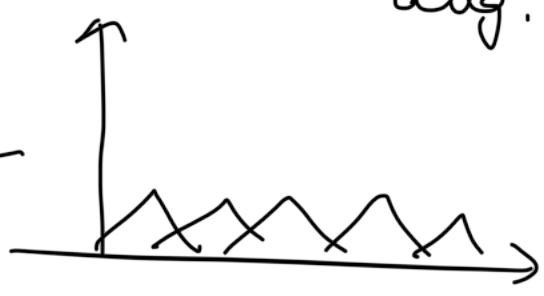


freq. domain DFT (FFT)

Triangular filter aug.



Log



it will be the feature vector $[\ ]_{80\times1}$ for the first 25 ms.

⇒ seq. of features, they follow time ordering

- GMM ⇒ often we assume diagonal (↓ uncorrelated) covar. matrices
- for speech, log mel filter bank
  ↓ DCT-2 (discrete cosine transform - 2)
  mel freq. cepstral coefficient

we retain only first 13 cepstral coeff.

delta
Δ ⇒ diff. btw. adjacent MFCC

$\Delta\Delta$ ⇒ diff. of diff. btw. adjacent MFCC.

↓

Concatenate to form the MFCC vectors.

L5 MFCC

1. Impt. of speech → speaker & listener
2. phonemems ⇒ spoken words
3. freq. as perceived mel

$$f_{mel} = 1125 \ln\left(1 + \frac{f_{hz}}{700}\right)$$

intensity as loudness

masking

4. processing the signal

features: mel. filter bank

↓

mel frequency cepstral coeff. (MFCC)

Excitaⁿ

$E(\Omega)$ ||||| → mouth filter → $s[n]$

↓ FT

$|S(\Omega)|$

$= V(\Omega) \times E(\Omega)$

- we use 25 ms, so small period of time,

that we assume that the speech signal will be stationary during this time.

- the seq. of mel filter bank features. Almost we 100 such $[]_{80\times1}$ frames for per second.

L2.6 Gaussian Review

$$f_x(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

(μ: mean, σ: SD)

↳ Gaussian distribu^n (bell-shaped curves)

- $(x-\mu)$ some kind of shift opera^n is happening.
- we use dataset to calculate μ & σ of gaussian distri. Generally, some estima^n techniques are used. use MLE (maximum likelihood estima^n)

$$\mu = \frac{1}{1000}\sum_{i=1}^{1000} x_i \qquad \sigma = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(x_i-\mu)^2}$$

- Likelihood

$$L(\mu, \sigma) = f_x(x_1 \cdots x_{1000}; \mu, \sigma)$$

$$\mu^*, \sigma^* = \underset{\mu,\sigma}{\operatorname{argmax}}\left(L(\mu, \sigma)\right)$$

↳ MLE*

$$= \underset{\mu,\sigma}{\operatorname{argmax}} \prod_{i=1}^{1000} f_x(x_i; \mu, \sigma)$$

$$= \underset{\mu,\sigma}{\operatorname{argmax}} \sum_{i=1}^{1000} \log\left(f_x(x_i; \mu, \sigma)\right)$$

- we can also use the gaussian model for classification problem.