EC864

SPEECH AND AUDIO PROCESSING

MANVITH PRABHU
21IEC228

SPEECH

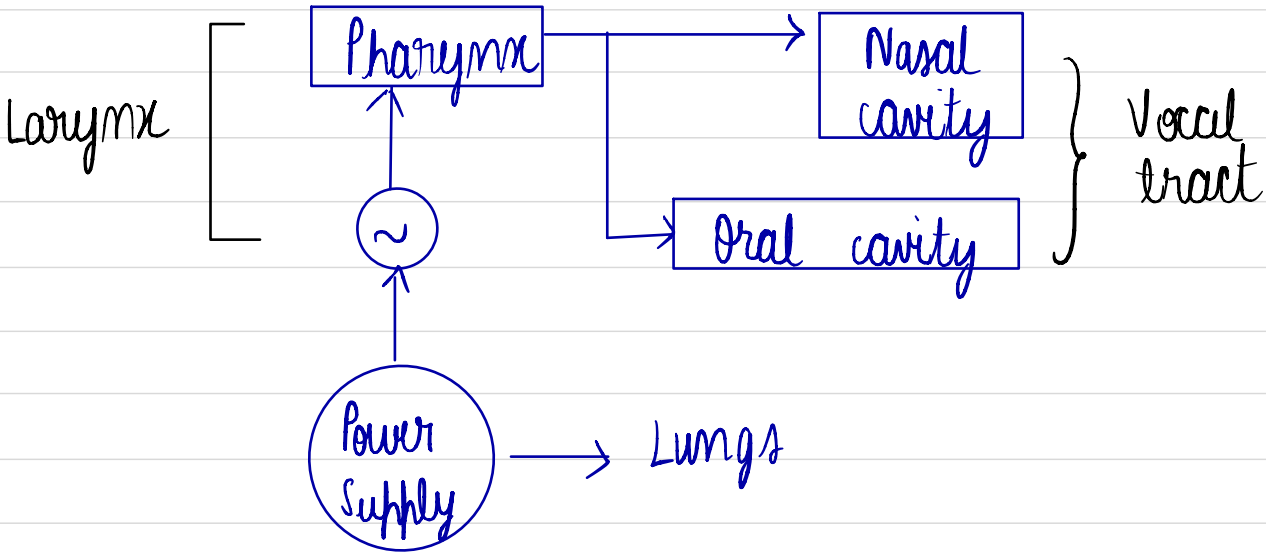
Speech generation in Humans:

Brain → Lungs → Vocal folds

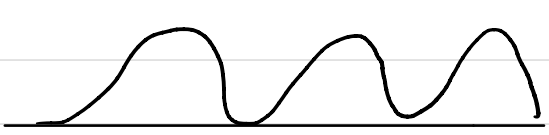
↓

Pressure signal ← Vocal tract

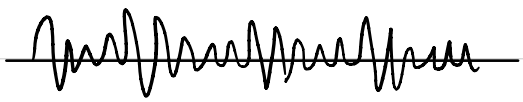
Block diagram



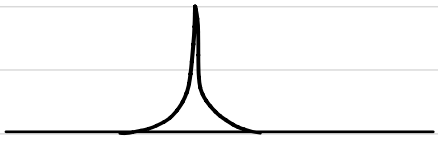
Types of sounds



→ Puff sound

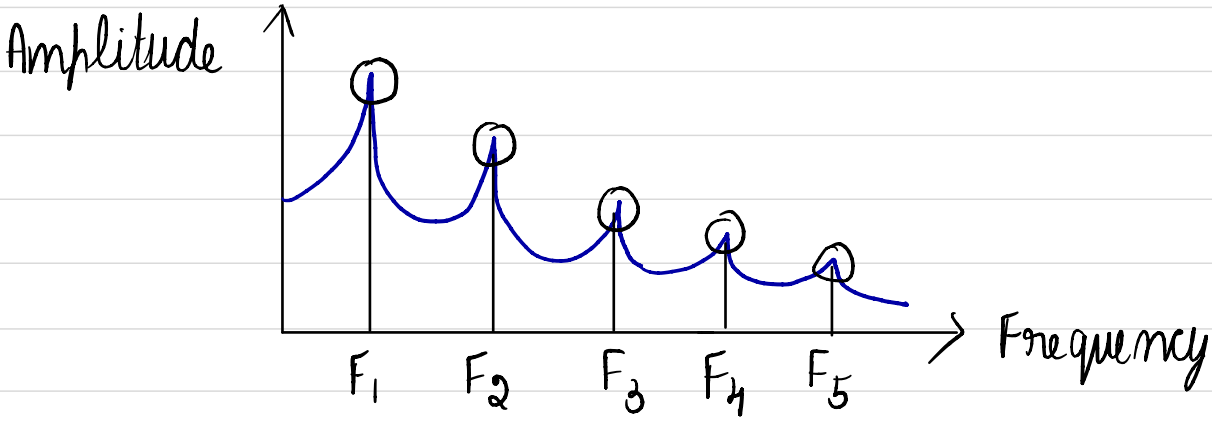


→ Noise



→ Impulse

NOTE: ECG is used to measure resistance between vocal folds → gives a waveform.



The peaks in frequency determine the vowel

→ On an average the frequency of voice of of men is of range 60 - 100 Hz
→ For women → 90 - 220 Hz
→ For children → 250 - 500 Hz
(Fundamental frequency)

Fundamental frequency

Pitch / Fundamental Period : The time duration of one glottal cycle as the pitch period.

$$\text{Fundamental frequency} = \frac{1}{\text{Pitch period}}$$

NOTE: Pitch is often used to describe the subjectively perceived "height" of a complex musical sound even when no single fundamental frequency exists.

Glottal Source based characteristics

Jitter: The variation in the pitch period across different glottal cycles

Shimmer: The amplitude variation in airflow velocity across glottal cycles.

Vocal Tract (VT)

→ Comprises oral cavity (from larynx to lips) and nasal passage

→ Average length of oral tract (adult male): ~17 cm.

→ For adult females, it is comparitively shorter

→ The pressure signal at the output of vocal folds (during voicing): time-varying buzz like sound.

→ VT: Spectrally "colour" the source which is important for making perceptually distinct speech sounds.

Spectral shaping (formant frequency)

- Under certain conditions, the relation between glottal airflow velocity input and vocal tract velocity output can be approximated by a linear filter with resonances.

- In speech science, these resonances known as formant frequencies / formants

- Formants change with different vocal tract configurations.

Categorisation of sound of Source

1) Voiced sounds: Speech sounds generated with a quasi-periodic glottal source.

2) Fricatives / unvoiced sounds: These sounds are generated by forming a constriction at some point in the vocal tract, (usually towards the mouth end) and forcing air through the constriction at a high enough velocity to produce turbulence.

- This creates a broad-spectrum noise source to excite the vocal tract.
  Eg: "sh" in should.

3> Plosive sounds: Generated from complete closure (usually towards the front of the VT), building up pressure behind the closure.
Eg: "ch" in chase

Types of Phonemes

→ Every language has a set of distinctive sounds or phonemes.
Eg: American english has around 42 phonemes.
1> Vowels
2> Diphthongs
3> Semivowels
4> Consonants.

• Each of these classes can be further subdivided

1> Vowels: These are produced by exciting a fixed VT with quasi-periodic pulses of air cause by vibration of vocal chords

2> Diphthongs: It is a gliding monosyllabic speech item that starts at or near the articulatory position of one vowel and moves towards the position for another.
Eg: /eI/ (as in bay)
/aI/ (as in boy)

/oI/ (as in boy) /oU/ (as in boat)
/jU/ (as in you) /aU/ (as in how)

3) Semivowels:
→ The group of sounds consisting of /w/, /l/, /r/, and /y/ are called semivowels because of their vowel like nature.
→ Generally characterized by a gliding transition in vocal tract area function between adjacent phonemes
→ Acoustic characteristics of these sounds are strongly influenced by the context in which they occur.
→ These sounds can be described as transitional, vowel-like sounds, and hence are similar in nature to the vowels and diphthongs.
Eg: "w" in we

4) Nasals: • Nasal consonants /m/, /n/ and /ng/ are produced with glottal excitation and vocal tract totally constricted at some point along oral passage.
• The velum is lowered so that air flows through the nasal tract, with sound being radiated at the nostrils.

Categorisation of speech sounds

- Phonemes arise from a combination of vocal fold & vocal tract articulatory features

- Articulatory features include:
  1> Vocal folds are vibrating or open
  2> The tongue position, height & it's constriction
  3> Velum state (nasal sound or not)

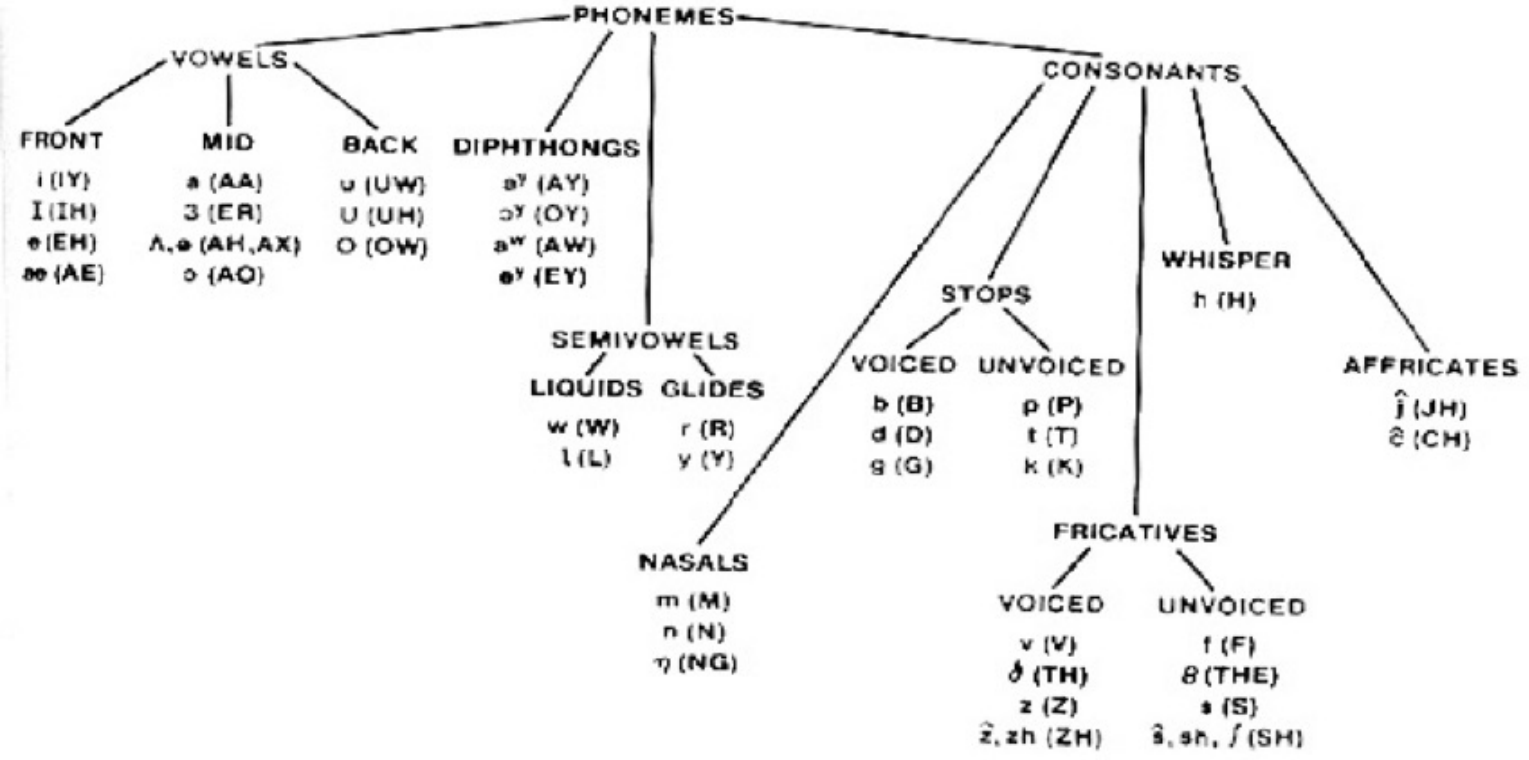
Vowels:

VT shape = f (tongue, jaws, lips & velum)

For non-nasalized vowels, the naval passage is not occupied with oral cavity

Tongue: Primary determinant of VT shape

Place of articulation: front, center & back of oral cavity



Consonants :

1> Nasals: Quasi-periodic airflow puffs from the vibrating vocal folds.

System: The velum is lowered, oral tract being constricted totally, and the air flows mainly through nasal cavity.

Nasal sounds radiates at nostrils

Example: /m/ - oral tract constricts at lips

Nasals have lower frequency compared to vowels

2> Fricatives :

Types : voiced & unvoiced

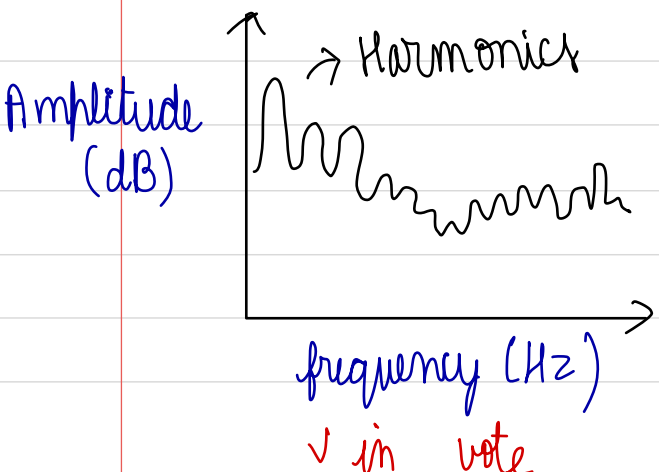
1> Source in unvoiced : Vocal folds are relaxed & not vibrating. Noise is generated by turbulent flow of air at some point of constriction (narrower than with vowels) along the oral tract.

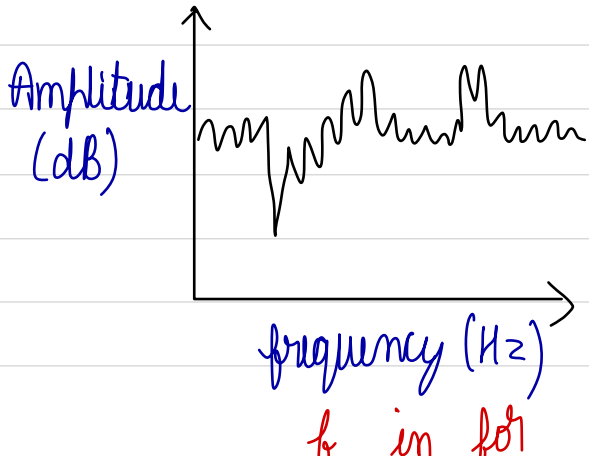
System : Location of the contriction by the tongue at the back, centre / front of the oral tract as well as teeth / lips.

Eg : f - (for) , h (he) , s (see)

2> Voiced : Source has a similar noise source and system characteristics to unvoiced fricatives with simultaneous
Eg : v in vote

Wideband spectrum : unvoiced : noisy spectrum
voiced : both noise and harmonics

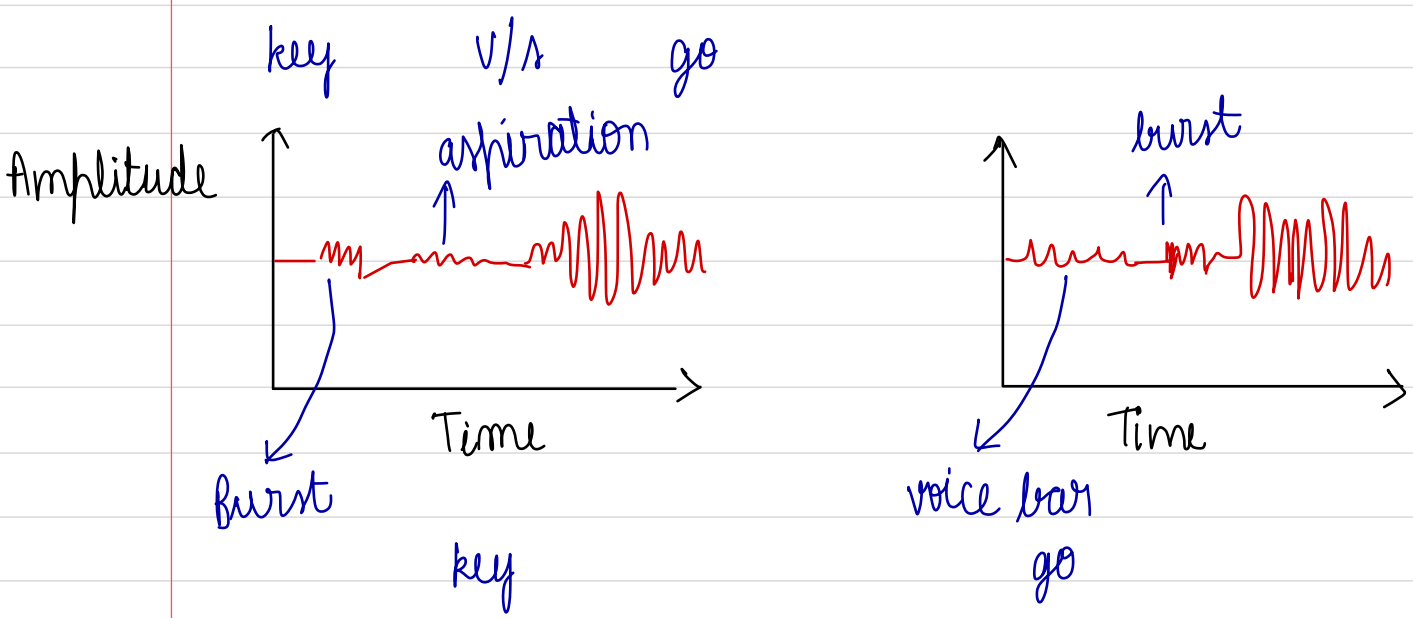




3> Plosive sounds

Unvoiced plosive : • A burst is generated at the release of the buildup of pressure behind a total constriction in the oral tract.
• This burst is idealized as an impulse source
• No vibration of vocal folds.
Eg : p (pay) , t (to) , k (key)

voiced plosives : Pressure is built behind an oral tract constriction, but vocal folds can also vibrate.
Presence of voice bar
Eg : b (be) , d (day) , g (go)



Diphthongs : They have 2 formants.

Continuant sounds: Eg: Vowels, Fricatives, Nasals
Produced by a fixed VT configuration excited by the appropriate source

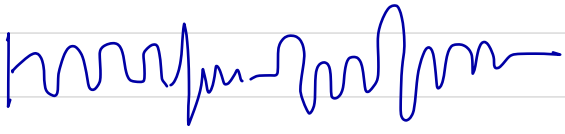
Non-continuant source: Eg: Diphthongs, Semivowels, stops, affricates.
Produced by a changing VT configuration excited by appropriate source.

Sound v/s vibration

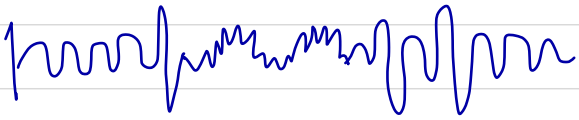
POA \ MOA	UVUA	UVA	VUA	VA
Palatal	ಕ, ಖ (K)	ಘ, ಘ [k^h]	ಗ, ಗ [g]	ಘ, ಘ [g^h]
Alveolar	ಟ, ಟ [ṭ]	ಠ, ಠ [$ṭ^h$]	ಡ, ಡ [ḍ]	ಢ, ಢ [$ḍ^h$]
Dental	ತ, ತ [t]	ಥ, ಥ [t^h]	ದ, ದ (d)	ಧ [d^h]
Bilabial	ಪ, ಪ [p]	ಫ, ಫ [p^h]	ಬ, ಬ [b]	ಭ, ಭ [b^h]

UVA:

VUA:



VA:

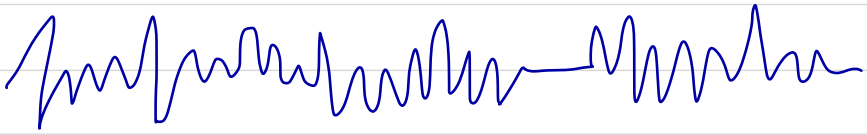


Stationary signal: A signal which can be mathema-tically defined.

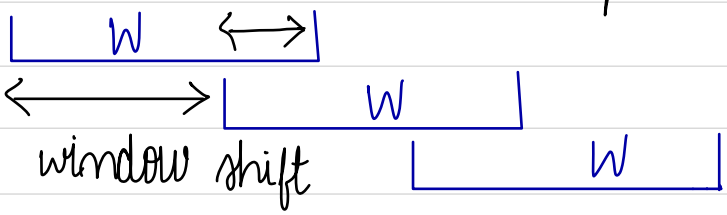
- Speech signal is non stationary and quasi periodic.
- So to analyze speech signal, we break speech signal to smaller chunks which is called window.

→ Windows can be non overlapping and overla-pping. Usually overlapping window size is preferred to avoid problems during reconstruction.

- If window size is W, usually window shift is usually W/2



window overlap.



Energy of stationary signals:

For a discrete signal $x[n]$, it's energy is given by: $E[n] = \sum_{n=-\infty}^{\infty} x^2[n]$

If frame size = N then energy of frame

$$E[n] = \sum_{m=n}^{n+N-1} x^2[m]$$

Eg: For a window of W = 25 ms and sampling frequency, F_s = 16 kHz. The no. of samples in the window is ?

Ans No of samples = $\frac{25 \times 16000}{1000}$ = 400 samples.

∴ Frame size, N = 400 samples.

Rectangular window

$$W_n(n) = \begin{cases} 1, & 0 \le n \le N \\ 0, & \text{otherwise} \end{cases}$$

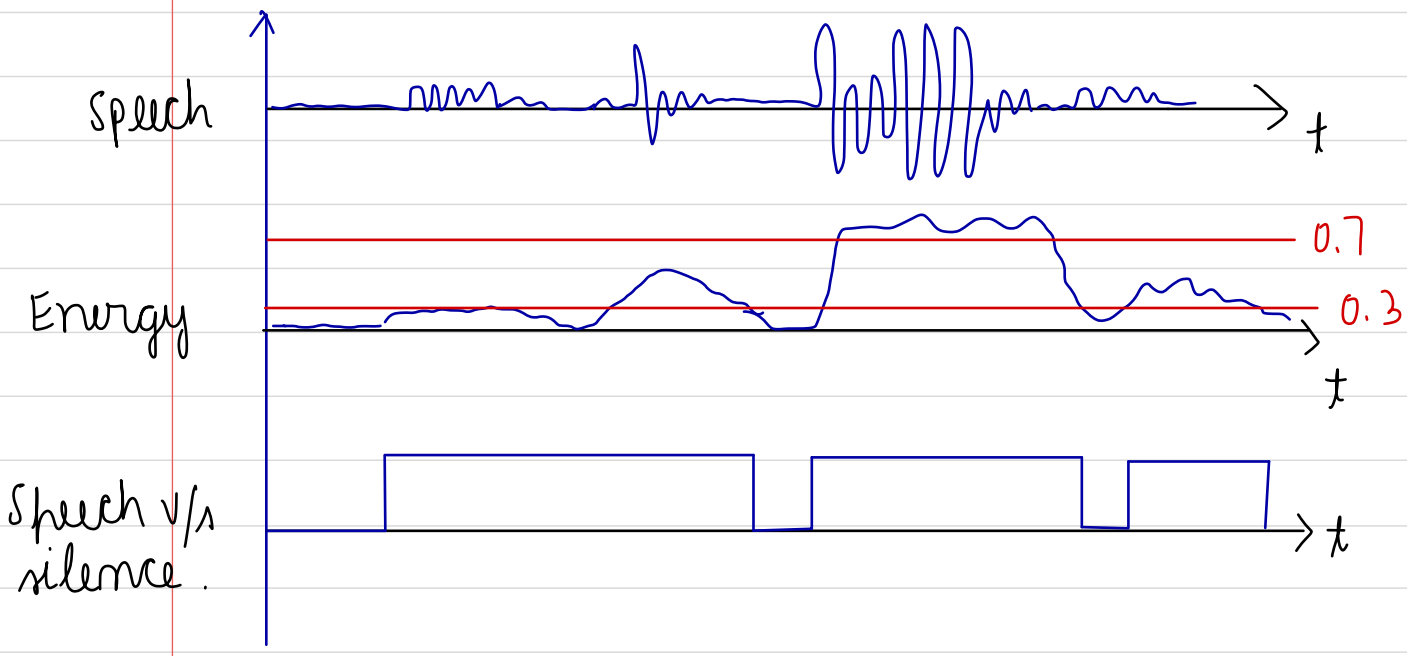
$$E[n] = \sum_{m=-\infty}^{\infty} \{ x[m] \cdot W[n-m] \}^2$$

For overlapping window $n = N/2$ shift is generally used.

NOTE: Energy of silence < Energy of unvoiced < Energy of voiced.

↓ Threshold 1 Eg (0.3)

↓ Threshold 2 Eg (0.7)

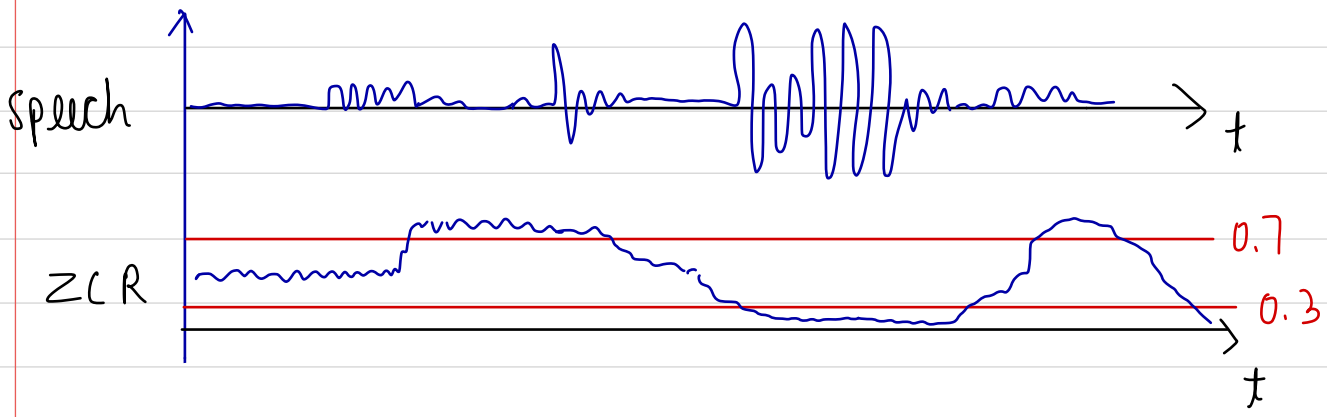


Zero crossing rate (ZCR)

$$ZCR = \frac{\text{No. of zero crossing}}{\text{Total duration}}$$

Zero crossing = No of sign changes.

Also $(ZCR)_{unvoiced} > (ZCR)_{silence} > (ZCR)_{voiced}$



NOTE: Signum function:

$$Sgn(x) = \begin{cases} 1 & x(n) > 0 \\ -1 & x(n) < 0 \end{cases}$$

For stationary signal:

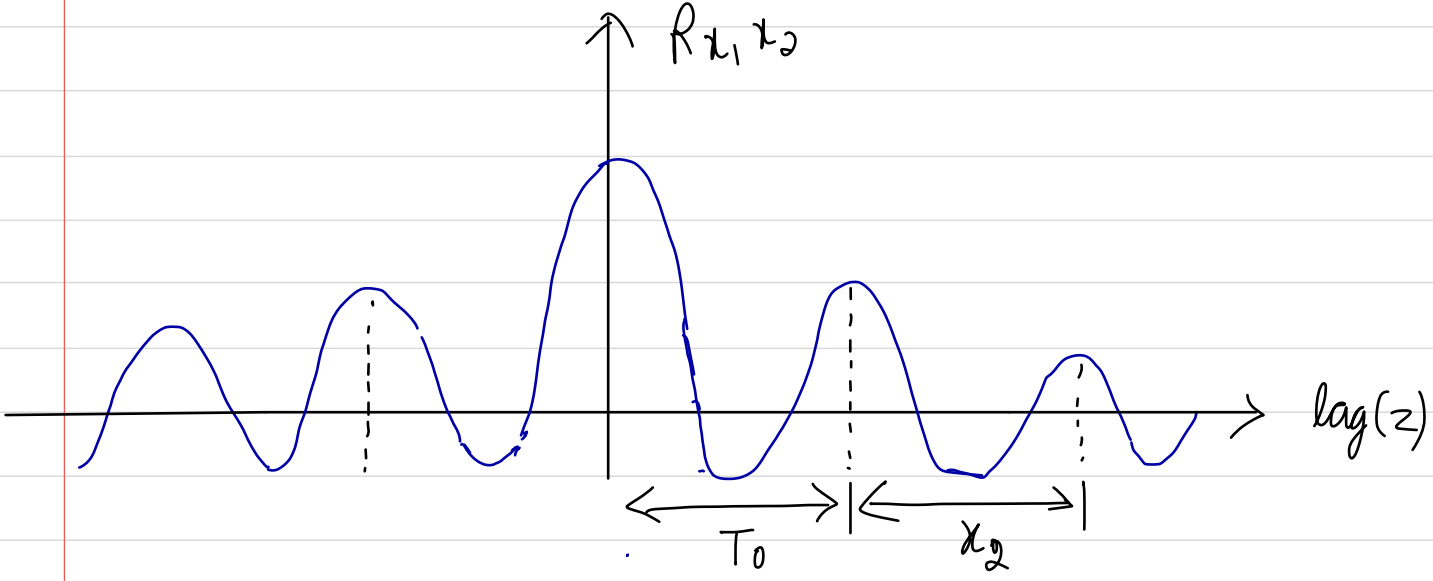
$$ZC = \sum_{m=-\infty}^{\infty} \frac{|sgn[x(m)] - sgn[x(m-1)]|}{2}$$

For total M samples:

$$ZCR = \frac{ZC}{M} = \frac{1}{M} \sum_{m=-\infty}^{\infty} \frac{|sgn[x(m)] - sgn[x(m-1)]|}{2}$$

For a windowed signal:

$$ZCR(n) = \sum_{m=n-N/2}^{n+N/2} \frac{1}{2} \Big| sgn[x(m)\cdot w(n-m)] - sgn[x(m-1)\cdot w(n-m+1)] \Big|$$



Convolution:

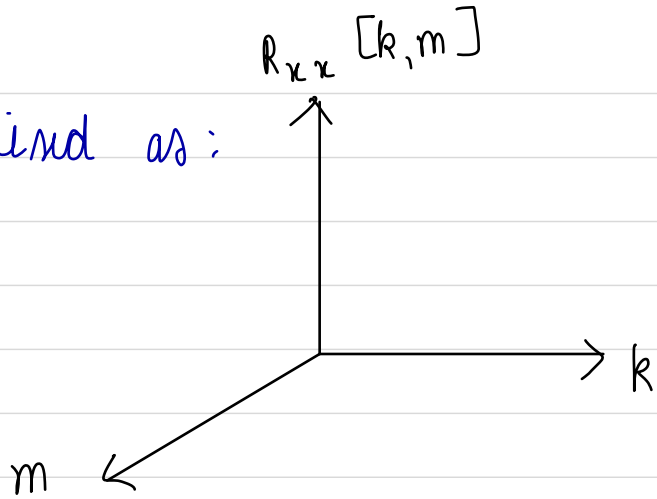
$$x_1[n] * x_2[n] = \sum_{m=-\infty}^{\infty} x_1[m]\, x_2[n-m]$$

Auto Correlation: $R_{xx}[k] = \sum_{n=-\infty}^{\infty} x[n]\, w[m-n] \cdot x[n+k]\, w[m-(n+k)]$

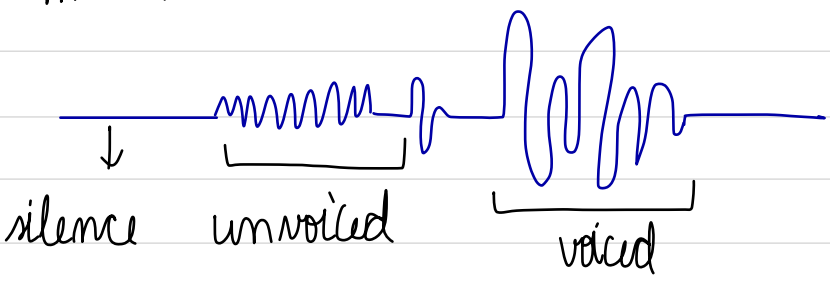
↓

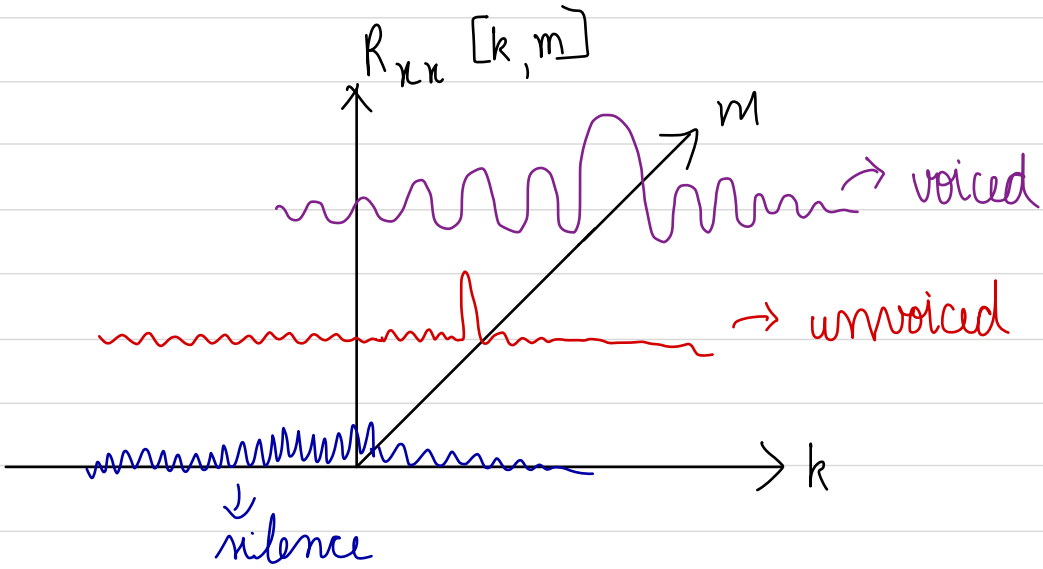
non stationary signal

This can be visualised as:



Consider signal:





Short time Fourier transform (STFT)

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n)\, e^{-j\omega n}$$

$$X(\omega, m) = \sum_{n=-\infty}^{\infty} x(n)\, w(m-n)\, e^{-j\omega n}$$

$$X(\omega_0, n) = \sum_{m=-\infty}^{\infty} x(m)\, w(n-m)\, e^{-j\omega_0 m} \quad (\omega \to \omega_0)$$

NOTE:

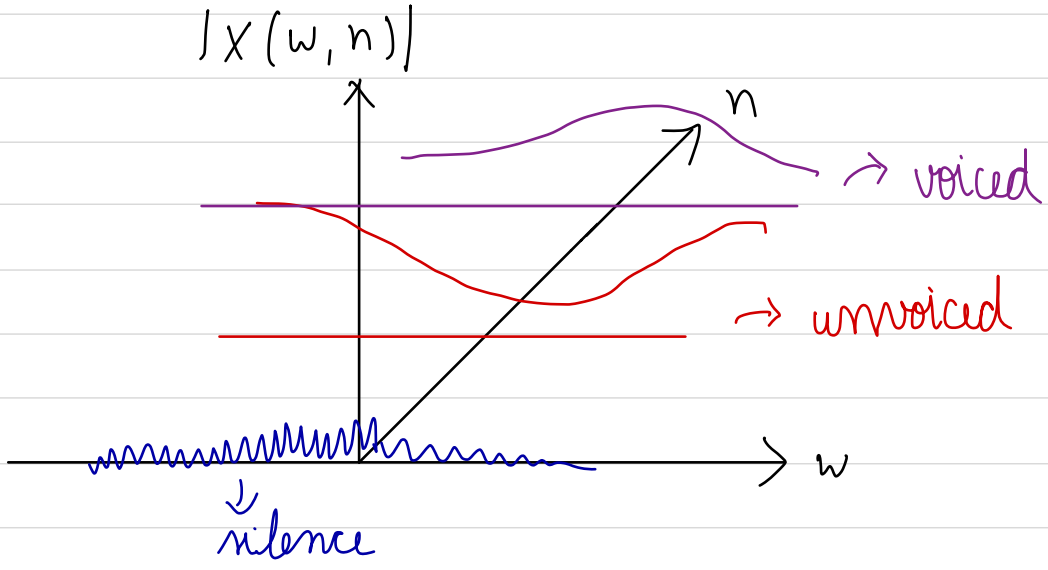
Let $p = n - m$, $p \to -\infty$ to ∞

$$X(\omega_0, n) = \sum_{p=-\infty}^{\infty} x(n-p)\, w(p)\, e^{-j\omega_0 (n-p)}$$

$$= e^{-j\omega n_0} \sum_{p=-\infty}^{\infty} x(n-p)\, \underbrace{w(p)\, e^{j\omega_0 p}}_{h(p)}$$

$$= e^{-j\omega n_0} \sum_{p=-\infty}^{\infty} x(n-p)\, h(p)$$

$h(p)$ is impulse response. So $w(p)\, e^{j\omega_0 p}$ acts as a impulse response.



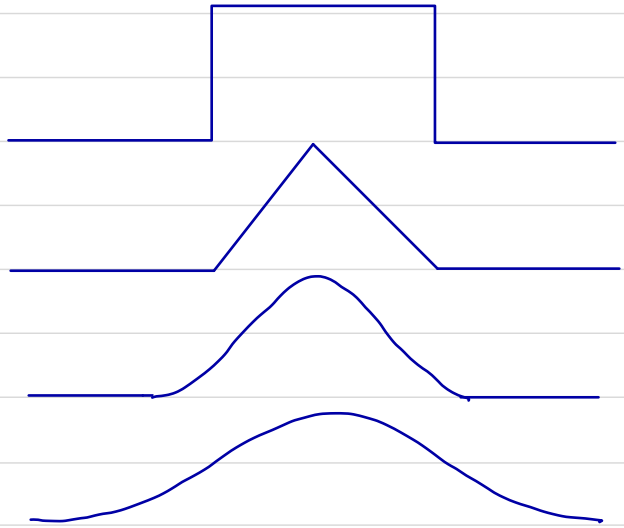
Types of windows

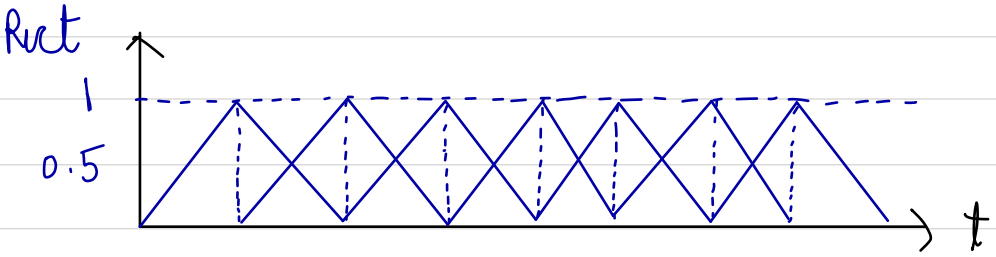
Rectangular

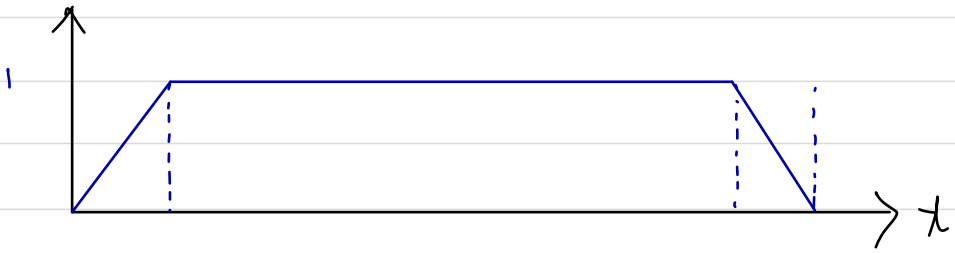
Triangular

Hamming

Hanning







← Reconstructed signal

STFT Spectral analysis

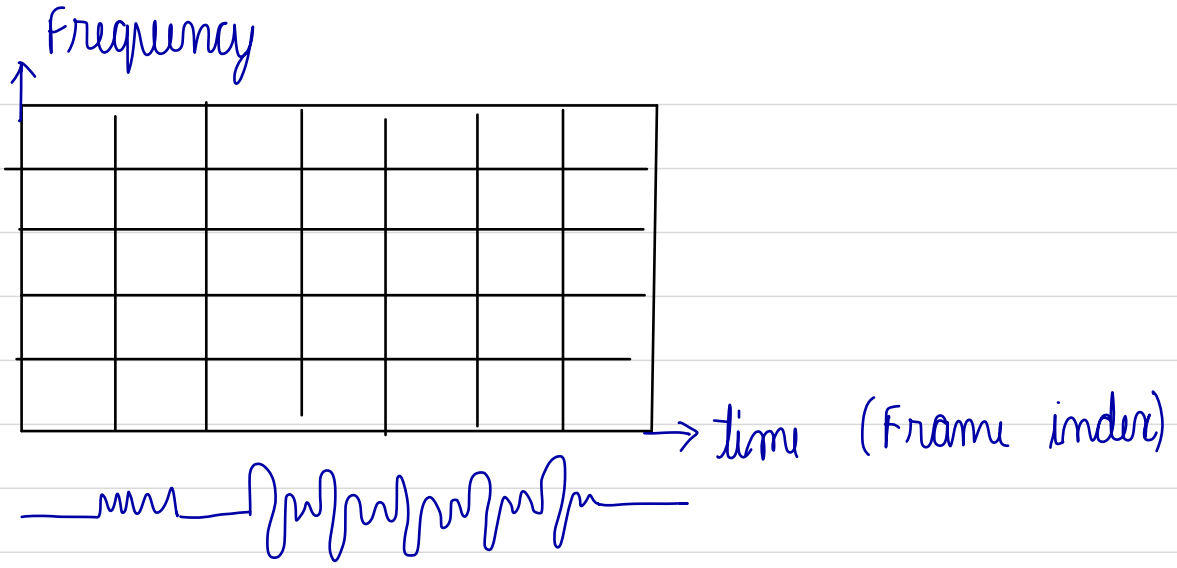
$x(t)$ —Sampling→ $x[n]$

Time domain

Aliasing in frequency domain

$x(t)$ —DTFT→ $X(\omega)$ —DFT→ $X(k)$

Aliasing in Time domain

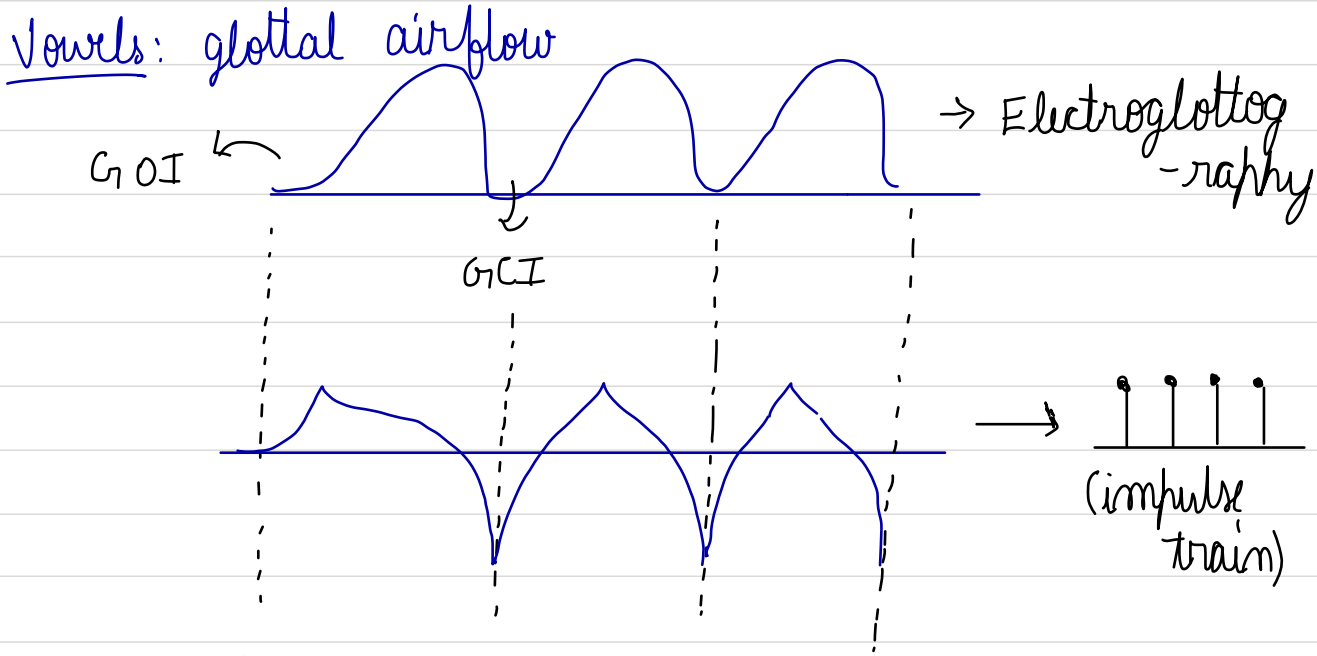


Excitation → Vocal tract → $S(n) = E(n) * h(n)$

$E(n)$ → $E(k)$

$h[n]$ → $H(k)$

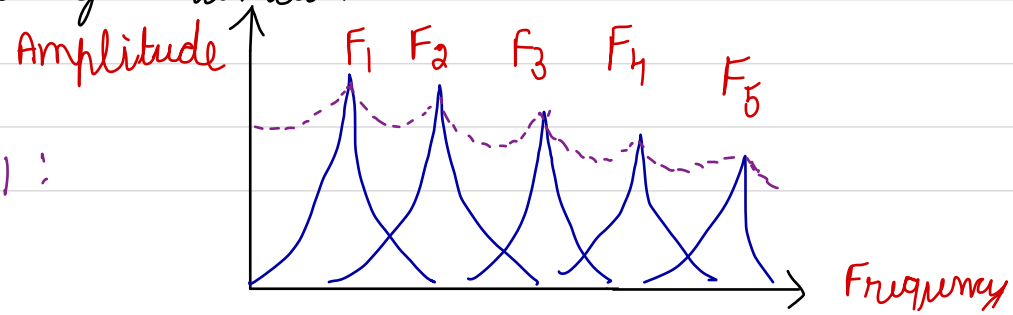
$S(k) = E(k) \cdot H(k)$



GOI = Glottal opening instance

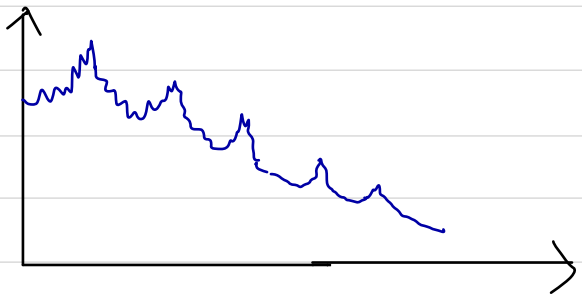
GCI = Glottal closing instance

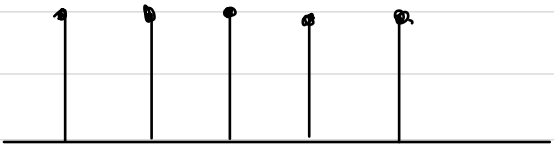
In frequency domain: →



g2:

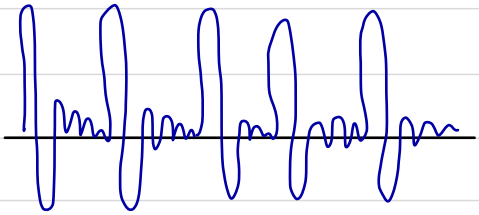
When g1 x g2 ;





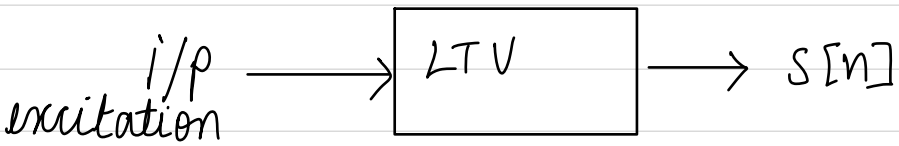
→ Impulse train

↓



Glottal pattern

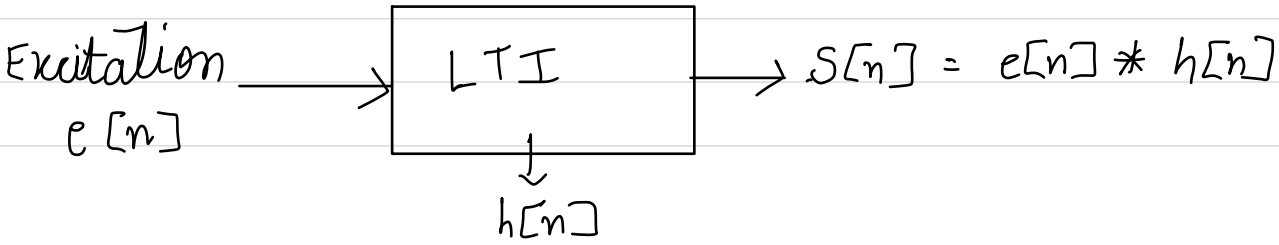
Homomorphic speech analysis



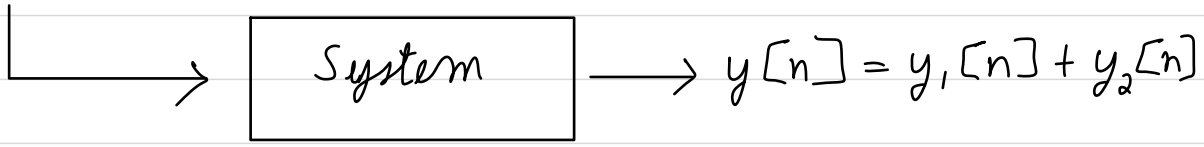
↳ slowly varies with time

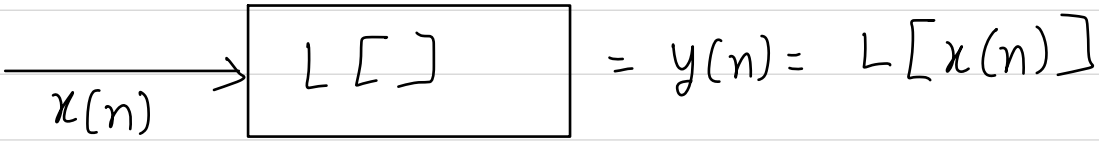
Linear time variant

Linear time invariant.



$x[n] = x_1[n] + x_2[n]$





↳ Linear operator

$$L(x[n]) = L(x_1[n] + x_2[n])$$

$$= L(x_1[n]) + L(x_2[n])$$

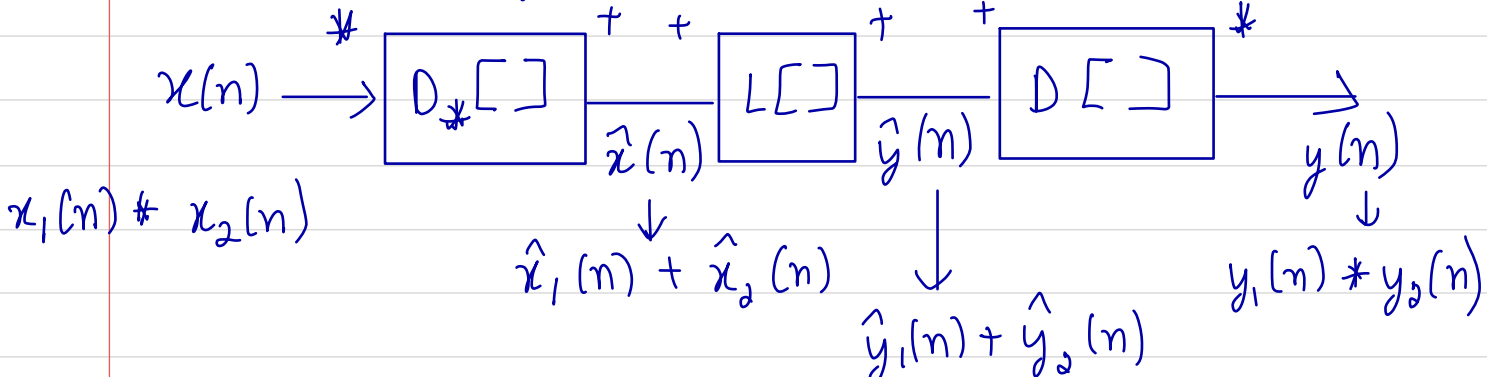
$$= y_1(n) + y_2(n)$$

$$y[n] = \sum_{k=-\infty}^{\infty} x(k)\, h(n-k) = h(n) * x(n)$$

$$H[x(n)] = H[x_1(n) * x_2(n)]$$

$$= H[x_1(n)] * H[x_2(n)]$$

$$= y_1(n) * y_2(n)$$



$$D_*[x(n)] = D_*[x_1(n) * x_2(n)]$$

$$= D_*[x_1(n)] + D_*[x_2(n)]$$

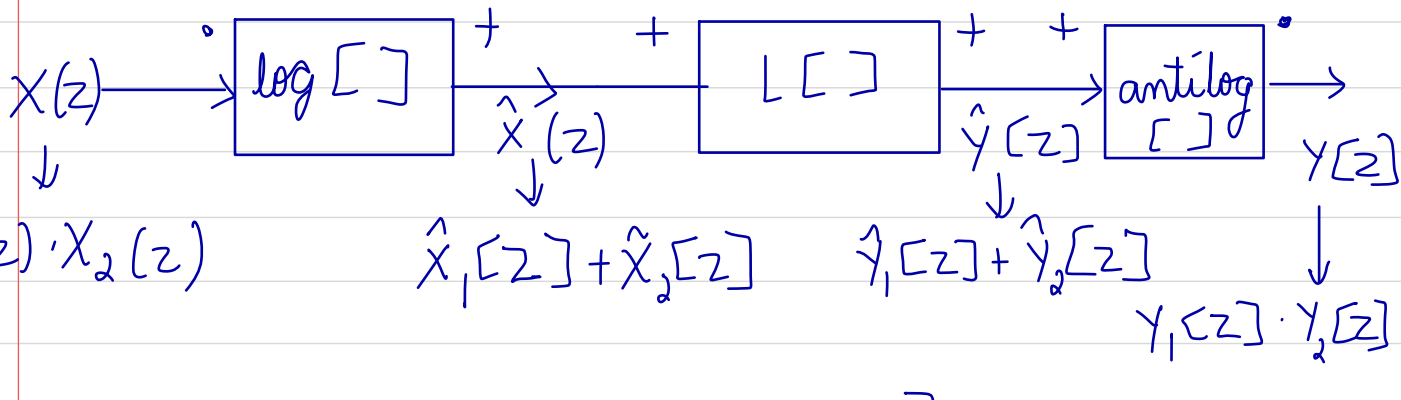
$$= \hat{x}_1(n) + \hat{x}_2(n)$$

eg: $a * b \xrightarrow{DFT} A[k] \cdot B[k]$

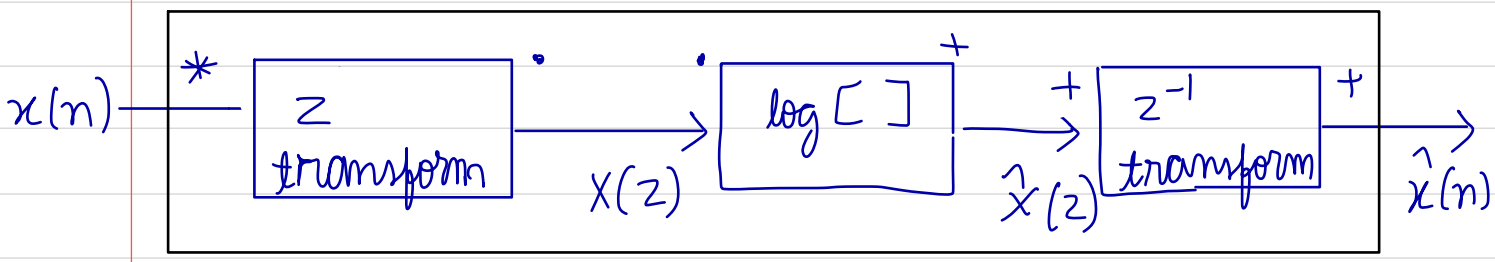
$\downarrow \log$

$\log(A[k]) + \log(B[k])$

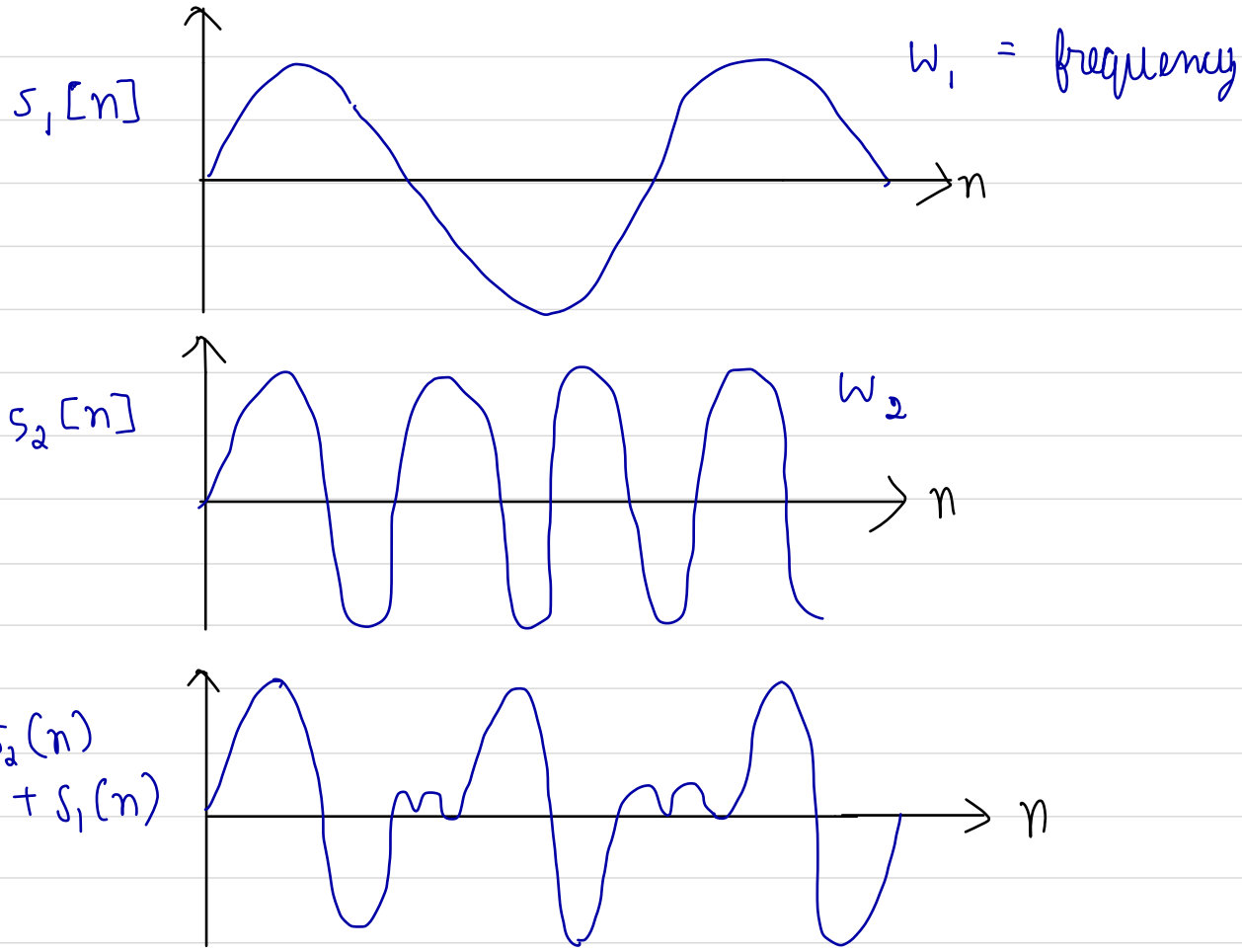
Frequency representation w.r.t convolution:



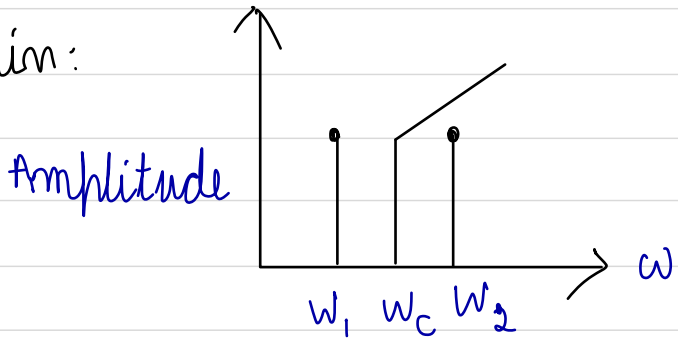
$$D_*^{-1}[\hat{y}(n)] = D_*^{-1}[\hat{y}_1(n) + \hat{y}_2(n)]$$



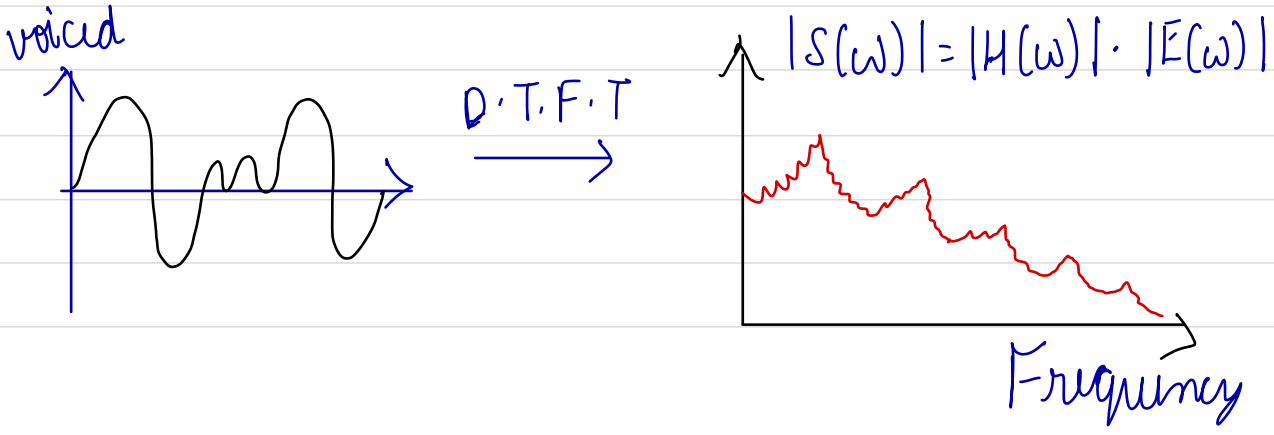
Homomorphic deconvolution



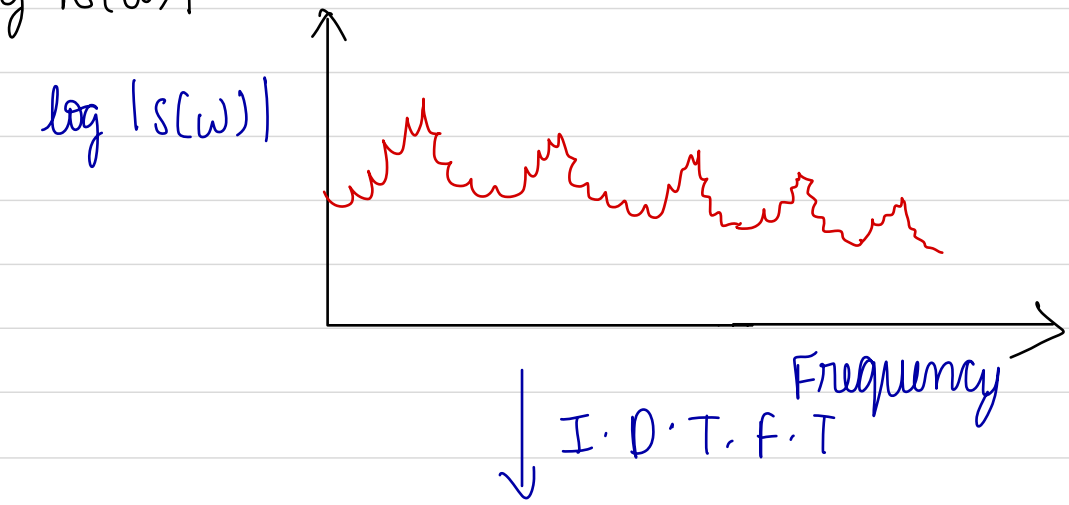
In frequency domain:



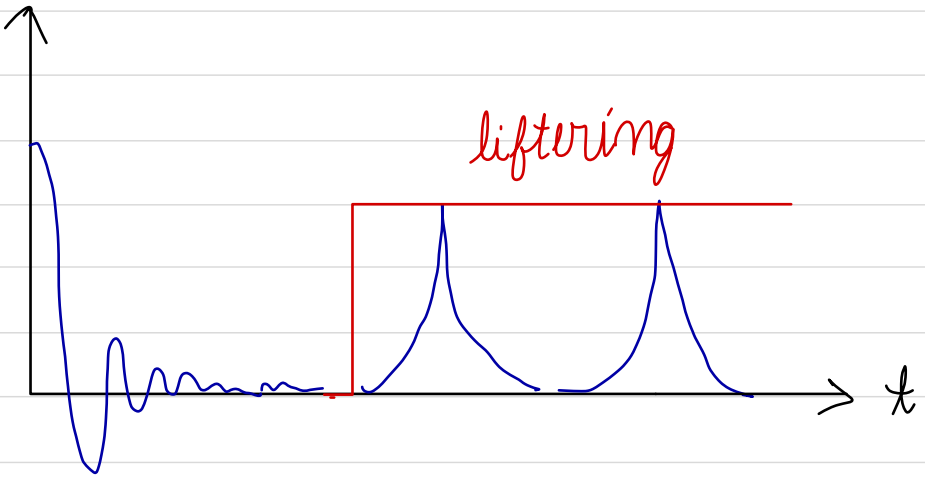
$$s[n] = e[n] * h[n]$$
$$S(w) = E[w] \cdot H[w]$$
$$= \log|E(w)| + \log|H(w)|$$



taking log $|S(\omega)|$



↓ I·D·T·F·T

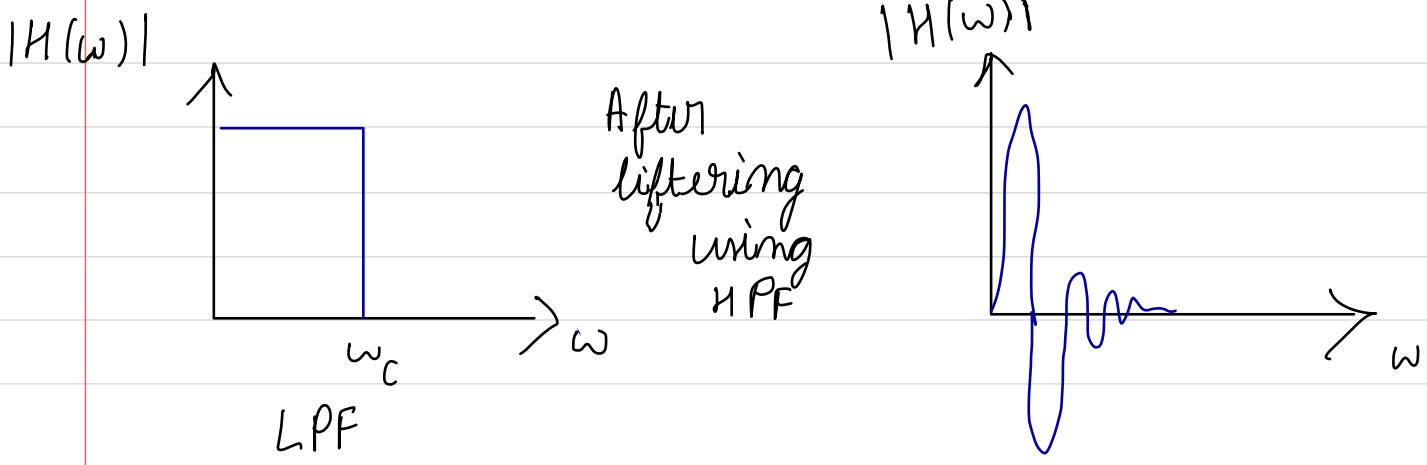


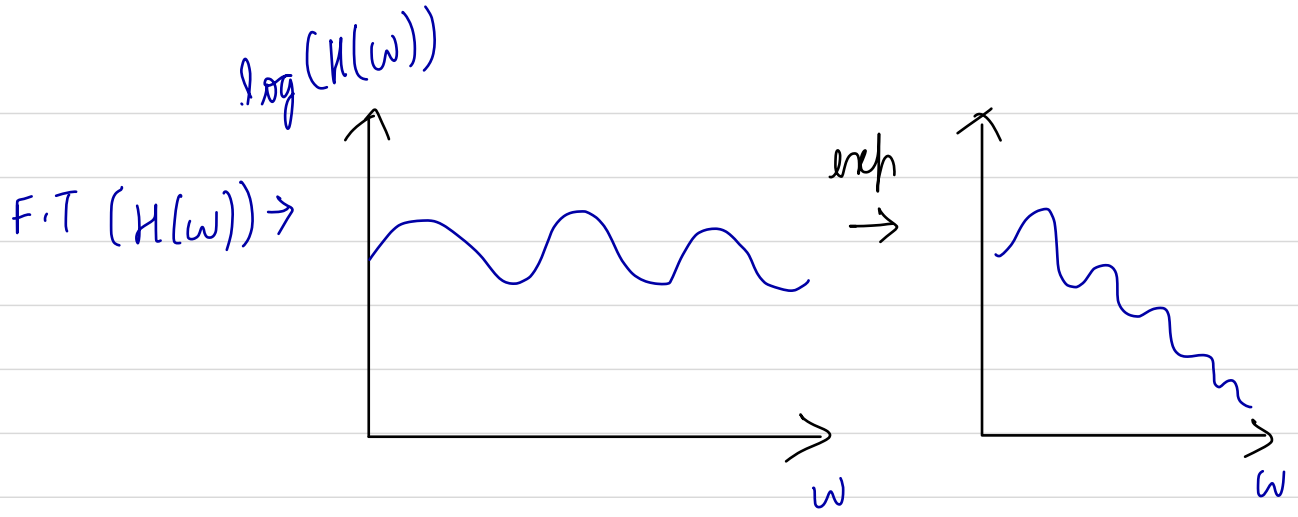
Cepstrum:

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S(\omega)| \cdot e^{j\omega n} \, d\omega$$

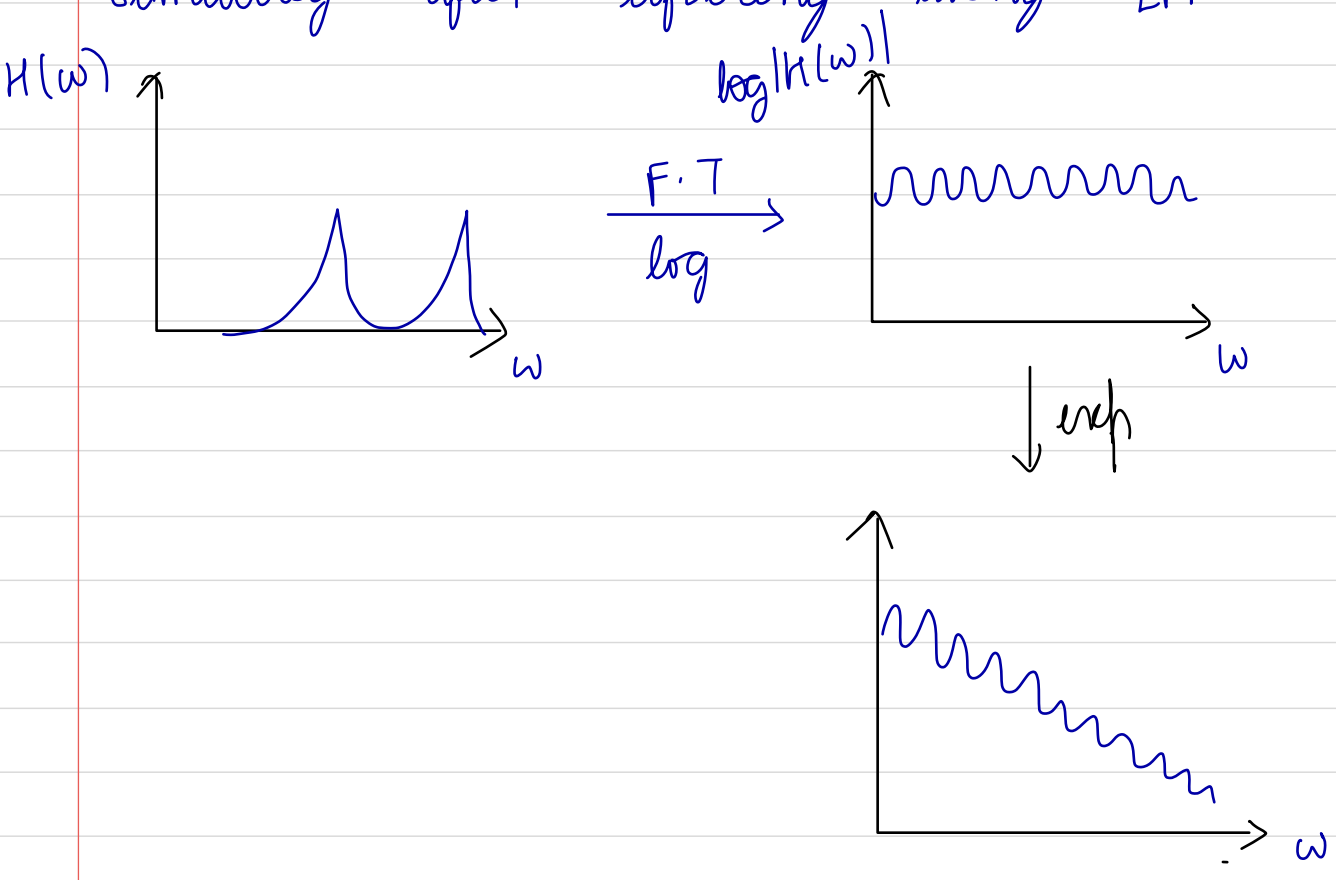
↓

Real part of cepstrum

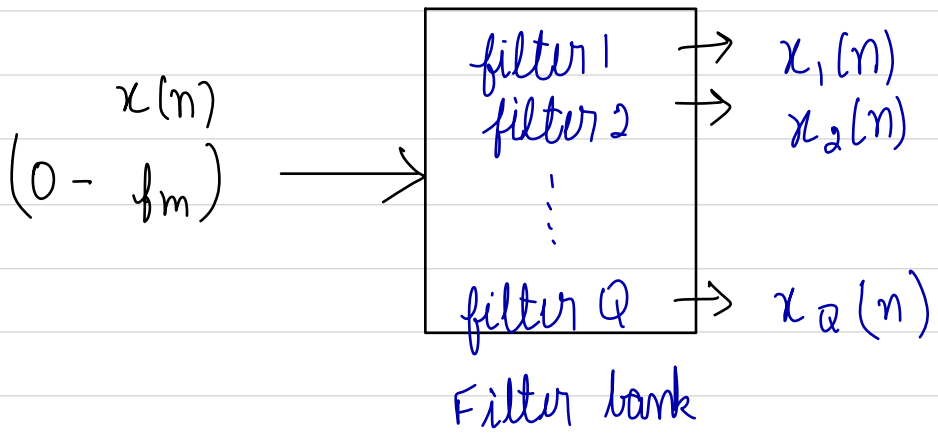




Similarly after liftering using LPF



Filter bank analysis



Filter bank

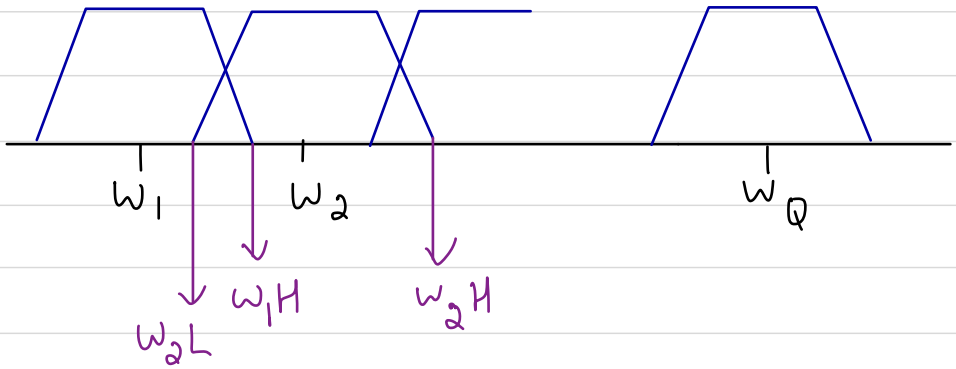
Liftering: Process of filtering out components outside the filter used. (choosing signal outside the filter)

Q = Number of bandpass filters.

$$Q = \frac{f_m}{B.W} \quad \text{or} \quad \frac{f_s}{2 \times BW}$$

Eg: if $f_s = 16000Hz$ & $BW = 100Hz$

$$Q = \frac{16000}{2 \times 100} = 80 //$$



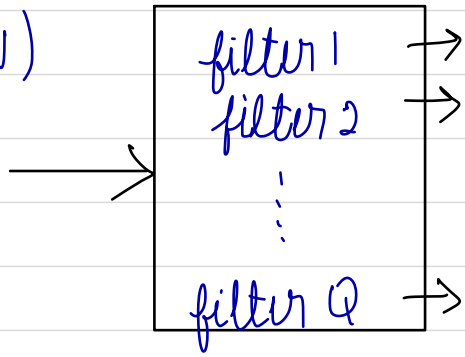
For non overlapping filters : $w_1H = w_2L$, $w_2H = w_3L$ & so on

NOTE: No. of frames = $1 + \frac{\text{len(signal)} - \text{len(frame size)}}{\text{len(hop size)}}$

$S_1(n) \Rightarrow (0-N)$

$\downarrow$ DFT

$|S_1(\omega)|$



$a_1 = (a_{11}, a_{12} \dots a_{1m})$

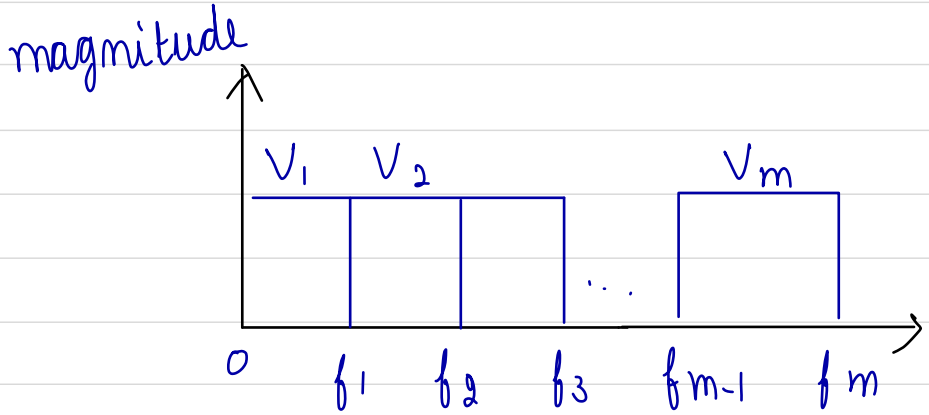
$a_2 = (a_{21}, a_{22} \dots a_{2m})$

a_Q

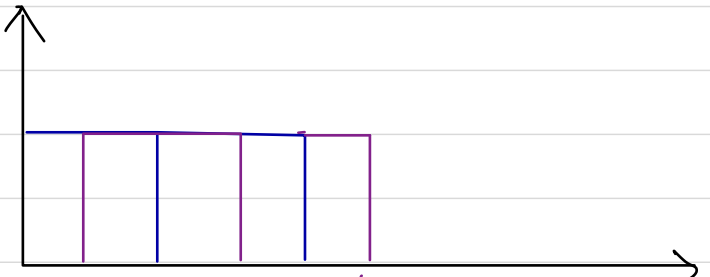
Filter bank

Energy: $\sum_{i=1}^{m} a_{1i}^2$, $\sum_{i=1}^{m} a_{2i}^2$, $\dots$, $\sum_{i=1}^{m} a_{Qi}^2$ $\Rightarrow$ $Q \times 1$

Uniform filter bank

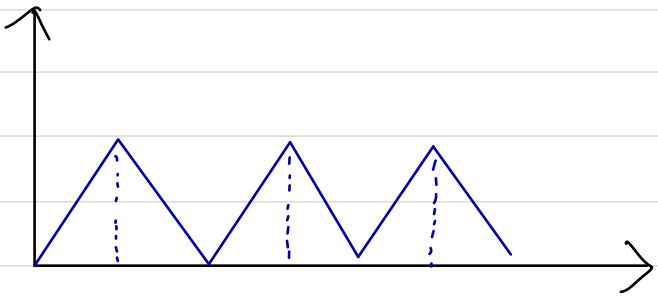


Mel frequency cepstrum

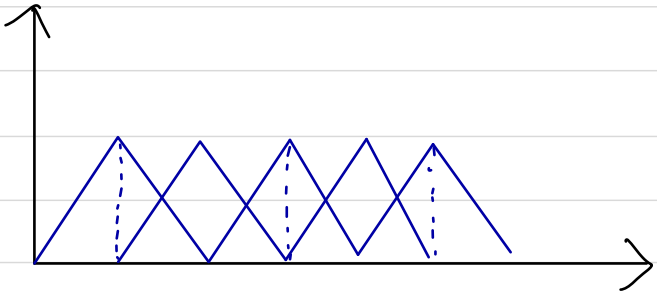


← uniform filter bank

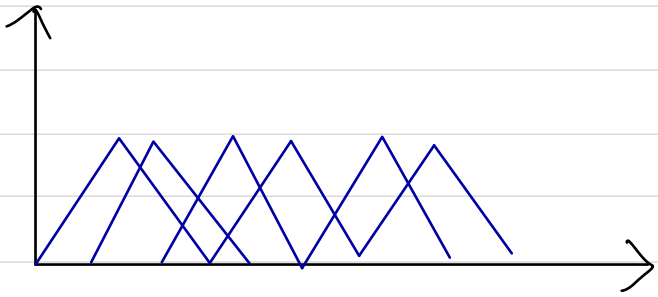
↳ 50% overlap



← Triangular filter bank (uniform, no overlap)

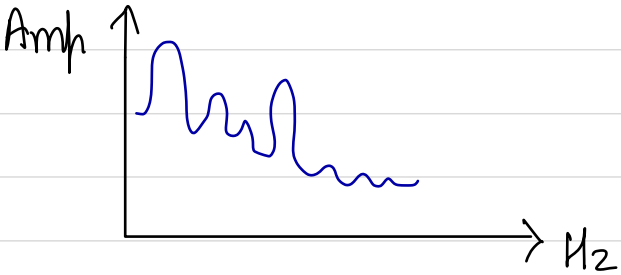


← Uniform 50% overlap Δ filter bank



← Non uniform Δ filter bank

$s[n] \xrightarrow{DFT} |s[k]|$



Hz ⇒ Physical frequency
Mel ⇒ Percieved frequency

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

or

$$f_{mel} = 2595 \log_{10} \left(1 + \frac{f_{Hz}}{700}\right)$$

Eg: If $f_{Hz} = 100$ Hz, find f_{mel}

$$f_{mel} = 1125 \log_e \left(1 + \frac{100}{700}\right)$$

$$= 150.22 //$$

NOTE.

f_{Hz}	f_{mel}
100	150.2
150	218.42
200	282.72
1500	1288.27
1550	1313.55
1600	1338.28

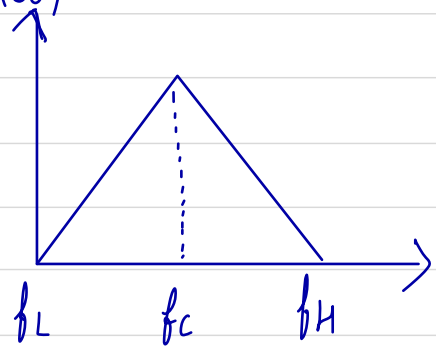
f_{mel}	f_{Hz}
100	64.95
150	99.65
200	135.92
1500	1949.31
1550	2069.49
1600	2195.13

Mel filter banks :

No. of filters = 26

Lower sensitivity for higher frequency.

Consider one filter: $H(\omega)$



$$H(\omega) = \begin{cases} \dfrac{f - f_L}{f_c - f_L}, & f_L \le f \le f_c \\ \dfrac{f_H - f}{f_H - f_c}, & f_c < f \le f_H \\ 0, & \text{otherwise} \end{cases}$$

Cepstrum analysis

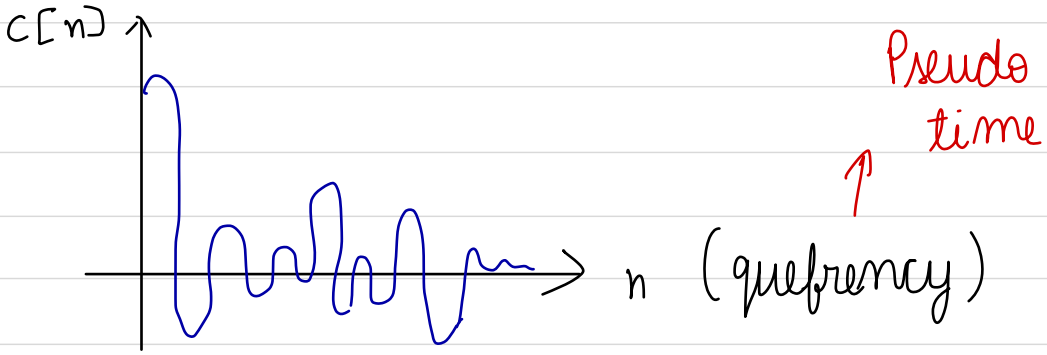
$s[n] = h[n] * e[n]$

↓ Homomorphic processing

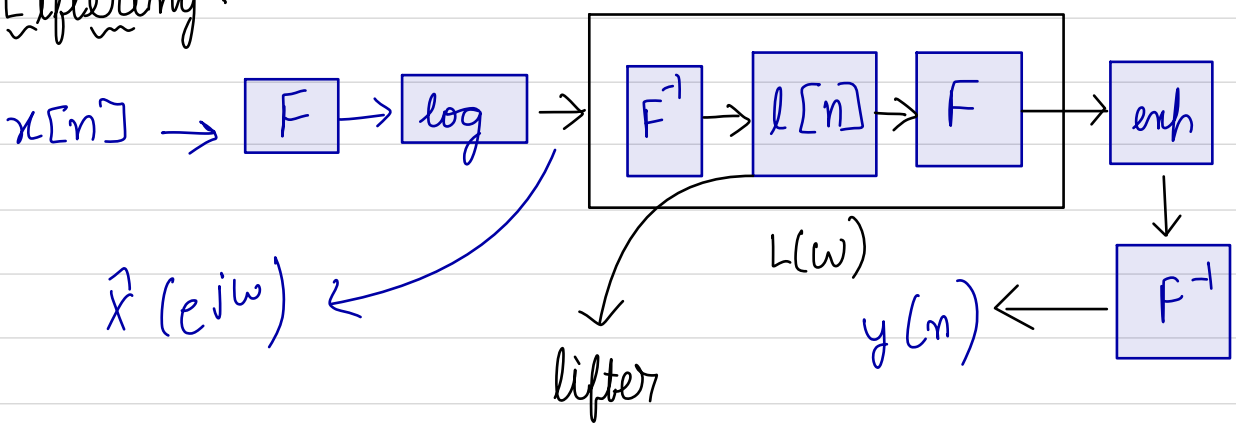
$\hat{h}[n] + \hat{e}[n]$

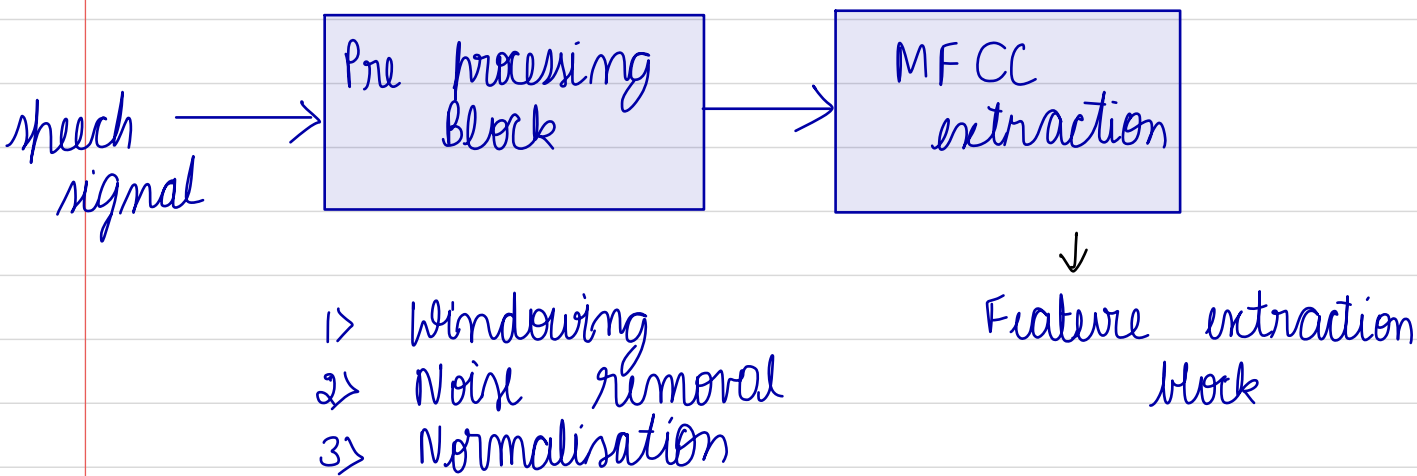
$$\hat{h}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \hat{X}(e^{j\omega}) e^{j\omega n} \, d\omega \quad \rightarrow \text{complex}$$

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |\hat{X}(e^{j\omega})| e^{j\omega n} \, d\omega \quad \rightarrow \text{real}$$



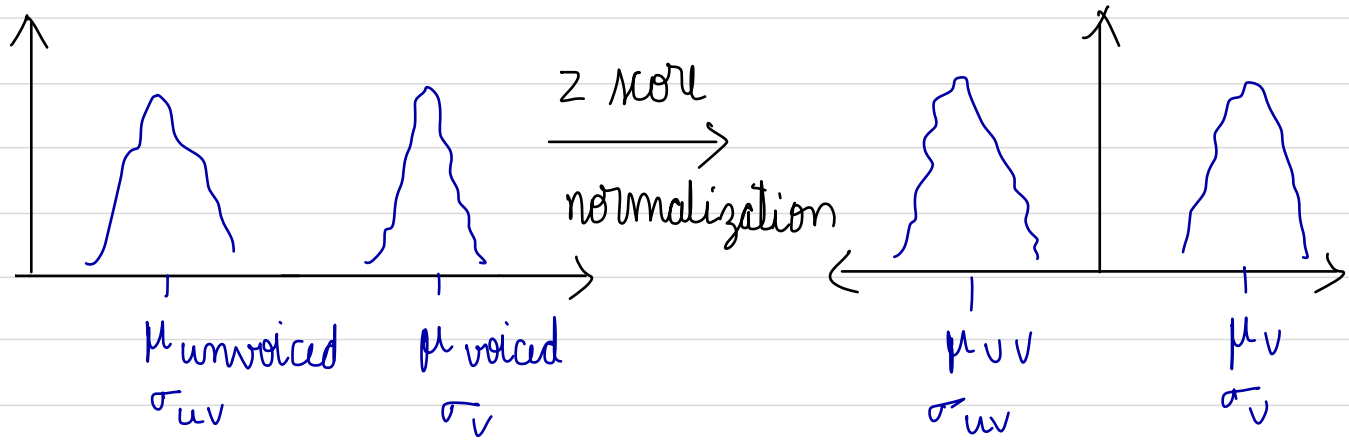
Liftering:

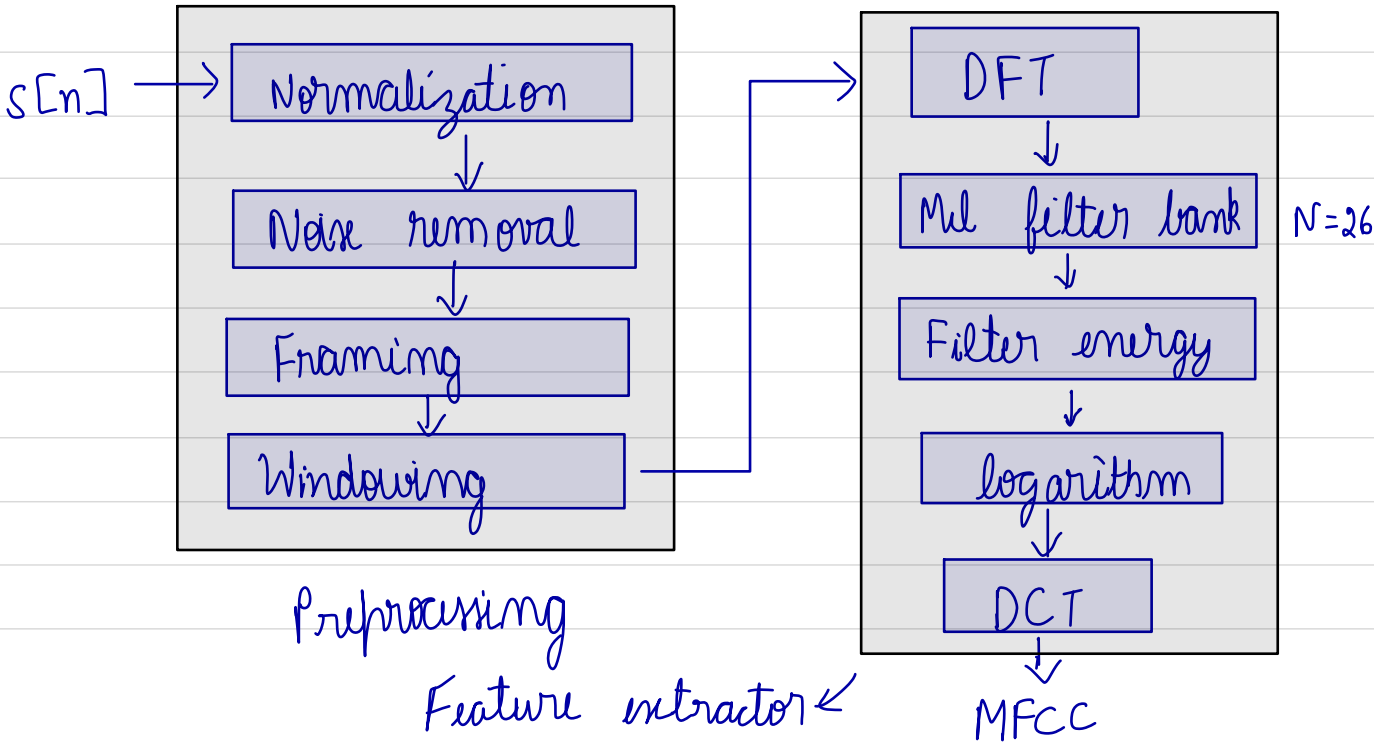




Z Score normalization:

$$z = \left(\frac{x-\mu}{\sigma}\right)$$





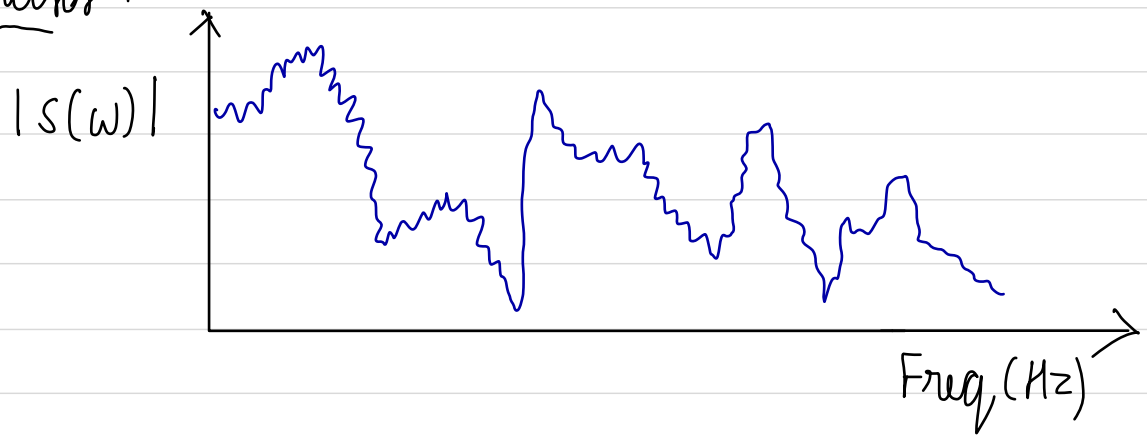
Windows: Hamming : $0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$

Hanning : $0.5\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right]$

MFCC delta: $mfcc_{i+1} - mfcc_{i-1}$

MFCC delta delta: $(mfcc_{i+1} - mfcc_i) - (mfcc_i - mfcc_{i-1})$

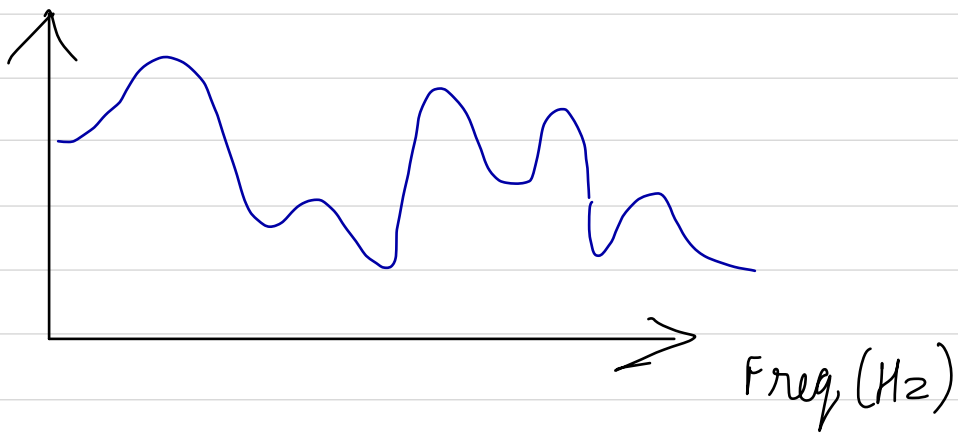
Preemphasis :



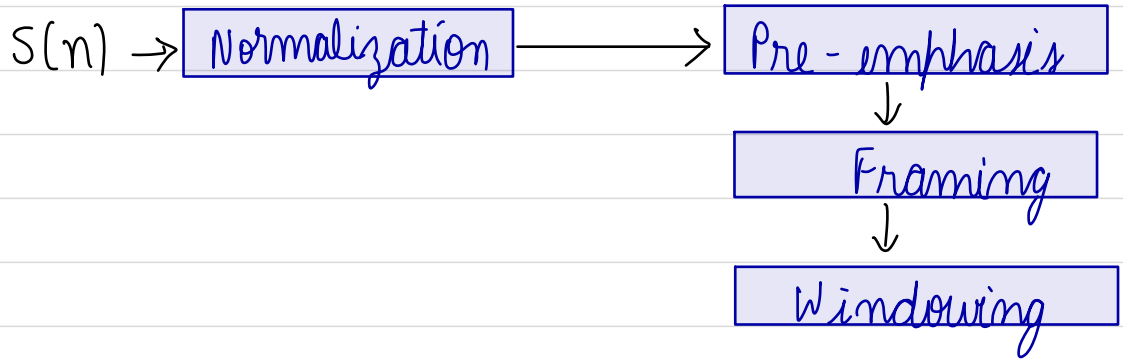
$P(z) = (1 - az^{-1})$ (acts as HPF)

$a = 0.68$ is mostly used for speech.

» Passing above signal through HPF



'a' decides amount of pre-emphasis



MFCCs

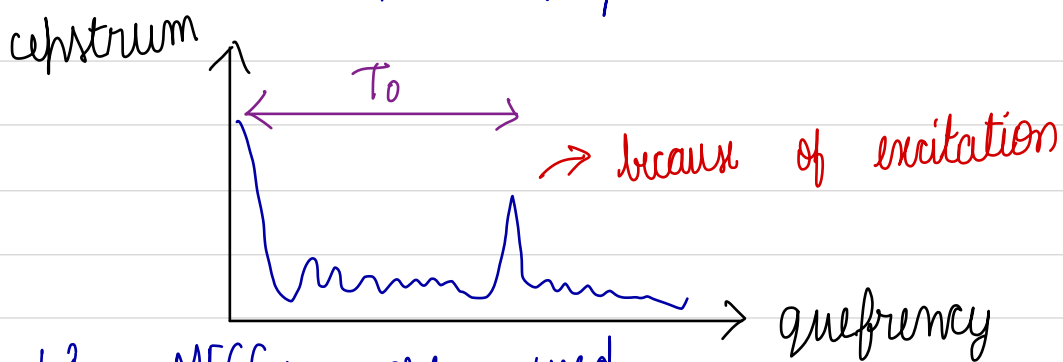
$$MFCC = [C_1, C_2, C_3, \ldots, C_N] \qquad N \leq 26$$

$$\Delta_i = C_{i+1} - C_{i-1}$$

$$\Delta = \frac{\sum_{n=1}^{N} n\,(C_{n+i} - C_{n-i})}{2\sum_{n=1}^{N} n^2}$$

$$\Delta\Delta_i = \Delta_{k+i} - \Delta_{k-i}$$

If N = 26 all will have shape 26×1



Usually 13 MFCCs are used

So 13×1 MFCC, 13×1 Δ, 13×1 ΔΔ

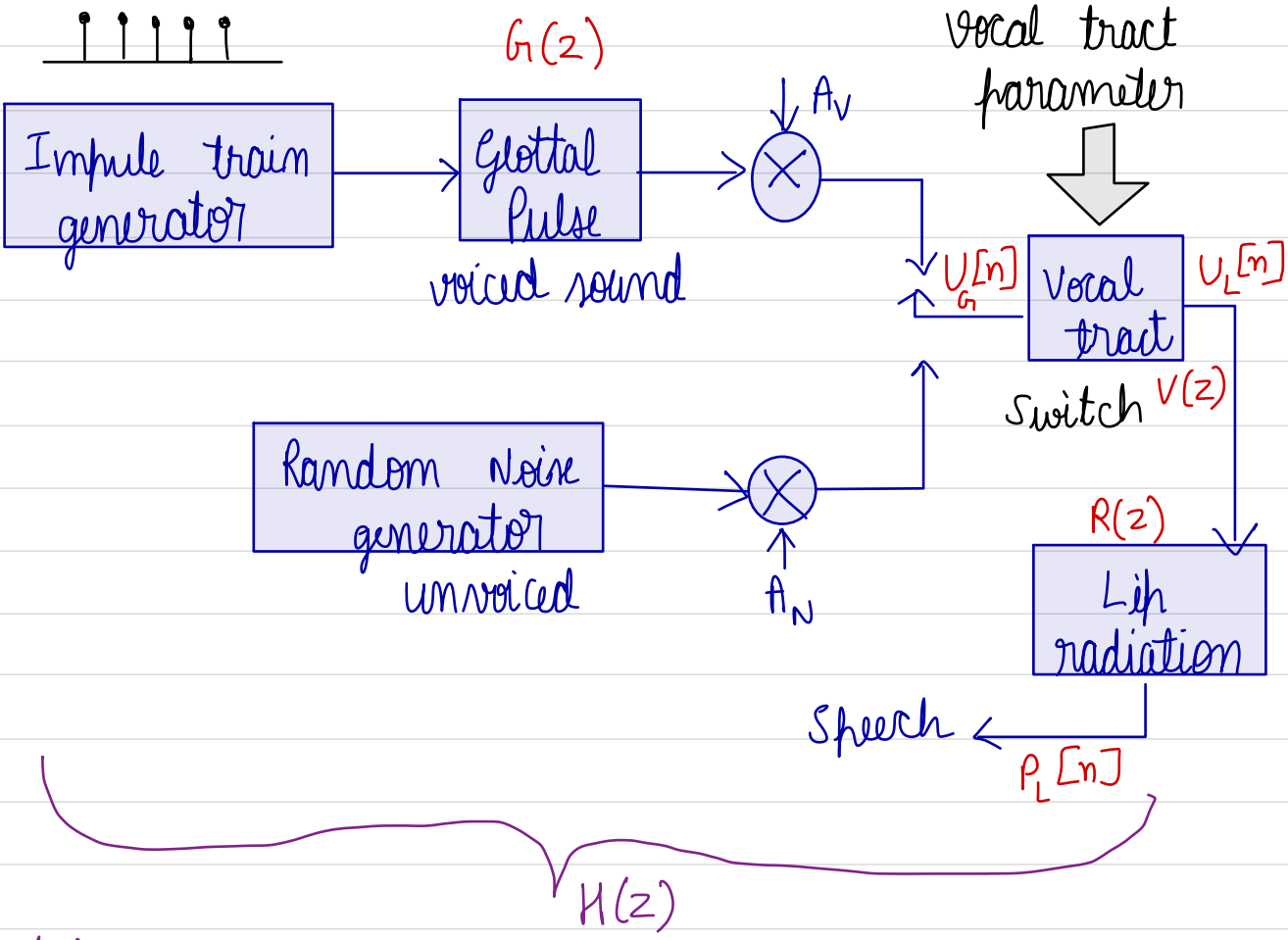
$MFCC^{T}_{full} = [MFCC, \Delta, \Delta\Delta]_{1\times 39}$

Fundamental frequency = $\frac{1}{T_0}$ (from MFCCs)

→ This method of finding fundamental frequency is more robust to noise than auto correlation method.

Linear Prediction coefficient : (LPC)

Speech Production Model :



$H(z) = G(z) \cdot V(z) \cdot R(z)$

where $G(z)$ = glottal transfer function

$V(z)$ = Vocal tract transfer function

$R(z)$ = Lip radiation transfer function

$$G(z) = \frac{1}{(1 - e^{-CT} z^{-1})^2} \qquad e^{-CT} \cong 1$$

$$= \left(\frac{z}{z-1}\right)^2$$

$$V(z) = \frac{G}{\prod_{k=1}^{N/2} (1 - 2\gamma_k \cos\theta_k z^{-1} + \gamma_k^2 z^{-2})^2}$$

$$R(z) = R_0 (1 - z^{-1})$$

$$H(z) = \frac{G R_0}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad \text{let } G R_0 = \sigma$$

$$= \frac{\sigma}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{S(z)}{U(z)}$$

Now $S(z) = H(z) \cdot U(z)$

$$S(z)\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right) = \sigma U(z)$$

$$S(z) = \sum_{k=1}^{p} a_k z^{-k} S(z) + \sigma U(z)$$

Taking inverse z transform

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + \sigma U(n)$$

when $U[n] = 0$

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] \Rightarrow a_1 s[n-1] + a_2 s[n-2] + \ldots$$
$$\ldots\ldots + a_p s[n-p]$$

↑ current sample

Here $\{a_1, a_2, a_3, \ldots, a_p\}$ are called LPC.

Prediction error:

$\tilde{s}[n]$ ↓ Predicted signal

$s[n]$ ↓ original signal

$$e[n] = s[n] - \tilde{s}[n]$$

↓ error

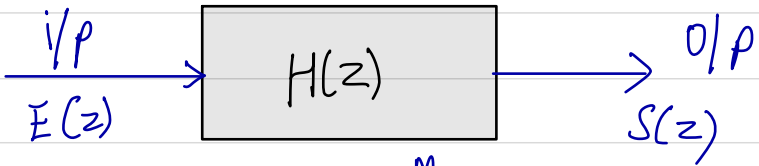
$e[n]$ is higher during transition from one region to another region.

$$e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k]$$

$$E(z) = S(z) - \sum_{k=1}^{p} a_k z^{-k} S(z)$$

$$= S(z)\left\{1 - \sum_{k=1}^{p} a_k z^{-k}\right\}$$

$$= S(z)\, A(z) \qquad \text{where} \quad A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$$



$$H(z) = \frac{S(z)}{E(z)} = \frac{\sum_{k=0}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}}$$

$$\sum_{k=0}^{N} a_k z^{-k} S(z) = \sum_{k=0}^{M} b_k E(z) z^{-k}$$

Taking inverse z transform:

$$\sum_{k=0}^{N} a_k s[n-k] = \sum_{k=0}^{M} b_k e[n-k]$$

$$a_0 s[n] + a_1 s[n-1] + \ldots + a_N s[n-N] = b_0 e[n] + b_1 e[n-1] + \ldots + b_M e[n-M]$$

Prediction : $$\tilde{s}[n] = -\sum_{k=1}^{p} a_k s[n-k]$$

Mean square error:

Instantaneous error:

$$e[n] = s[n] - \tilde{s}[n] = s[n] + \sum_{k=1}^{p} a_k s[n-k]$$

$$MSE = \sum_n e^2[n] = \sum_n \left(s[n] + \sum_{k=1}^{p} a_k s[n-k] \right)^2$$

For $k=1$, $\frac{\partial E}{\partial a_1} = 0$

$$\frac{\partial E}{\partial a_1} = \sum_n \left(s[n] s[n-1] + \sum_{k=1}^{p} a_k s[n-1] s[n-k] \right) = 0$$

for $k=2$,

$$\frac{\partial E}{\partial a_2} = \sum_n \left(s[n] s[n-2] + \sum_{k=1}^{p} a_k s[n-2] s[n-k] \right) = 0$$

Illy for $k = p$,

$$\frac{\partial E}{\partial a_k} = \sum_n \left(s[n]\, s[n-p] + \sum_{k=1}^{p} a_k s[n-p]\, s[n-k] \right) = 0$$

$$\Rightarrow \quad \sum_n s[n]\, s[n-p] = -\sum_{k=1}^{p} a_k \sum_n s[n-k]\, s[n-p]$$

Autocorrelation method:

$$R(i) = \sum_{n=-\infty}^{\infty} s[n]\, s[n+i]$$

Also $R(i) = R(-i)$

$$R(i) = \sum_{n=-\infty}^{\infty} s[n]\, s[n-i]$$

$$R(i) = -\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s[n-k]\, s[n-i]$$

where $i = 1, 2, \ldots, p$

$$R(i) = -\sum_{k=1}^{p} a_k R(i-k)$$

Total minimum error (E_p):

$$E = \sum_n \left(s[n] + \sum_{k=1}^{p} a_k s[n-k] \right)^2$$

$$= \sum_{n=-\infty}^{\infty} \left\{ s^2[n] + \left(\sum_{k=1}^{p} a_k s[n-k] \right)^2 + 2 s[n] \sum_{k=1}^{p} a_k s[n-k] \right\}$$

$$E = \underbrace{\sum_{n=-\infty}^{\infty} s^2[n]}_{\text{Term 1}} + \sum_{n=-\infty}^{\infty} \left(\sum_{k=1}^{p} a_k s[n-k] \right) \left(\sum_{i=1}^{p} a_i s[n-i] \right) + \underbrace{2 s[n] \sum_{k=1}^{p} a_k s[n-k]}_{\text{Term 2}}$$

$$E = \text{Term 1} + \sum_{i=1}^{p} a_i \sum_{n=-\infty}^{\infty} \sum_{k=1}^{p} a_k s[n-k]s[n-i] + 2\sum_{n=-\infty}^{\infty} \sum_{k=1}^{p} a_k s[n]s[n-k]$$

$$= \sum_{n=-\infty}^{\infty} s^2[n] - \sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s[n]s[n-k] + 2\sum_{n=-\infty}^{\infty} \sum_{k=1}^{p} a_k s[n]s[n-k]$$

$$= \sum_{n=-\infty}^{\infty} s^2[n] + \sum_{n=-\infty}^{\infty} \sum_{k=1}^{p} a_k s[n]s[n-k]$$

$$\therefore E_p = R(0) + \sum_{k=1}^{p} a_k R(k)$$

Levinson Durbin (LD) Recursive approach

$$R(0) + \sum_{k=1}^{p} a_k R(k) = E_p$$

$$R(i) + \sum_{k=1}^{p} a_k R(|i-k|) = 0 \quad ; \quad i = 1, 2, \ldots, p$$

$$R^{(i-1)} \; a^{(i-1)} = e^{(i-1)}$$

$R^{(i-1)}$: matrix; $a^{(i-1)}$: vector; $e^{(i-1)}$: vector

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(i-2) & R(i-1) \\ \vdots & \vdots & & \vdots & \vdots \\ R(i-1) & R(i-2) & & R(1) & R(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1^{(i-1)} \\ \vdots \\ a_{i-1}^{(i-1)} \end{bmatrix} = \begin{bmatrix} E_{i-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

i^{th} iteration autocorrelation matrix:

1 ←

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(i-2) & R(i-1) & R(i) \\ R(1) & R(0) & \cdots & & \vdots & \vdots \\ \vdots & \vdots & & & & \\ R(i-1) & R(i-2) & \cdots & R(0) & R(1) & \\ R(i) & R(i-1) & \cdots & R(1) & R(0) & \end{bmatrix} \begin{bmatrix} 1 \\ a_1^{(i-1)} \\ \vdots \\ a_{i-1}^{(i-1)} \\ 0 \end{bmatrix} = \begin{bmatrix} E_{i-1} \\ 0 \\ \vdots \\ 0 \\ \gamma^{i-1} \end{bmatrix}$$

$$\gamma^{(i-1)} = R(i) - \sum_{j=1}^{j-1} a_{(i-j)}^{(i-1)} R(j)$$

2 ←

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(i) \\ \vdots & \vdots & & \vdots \\ R(i) & R(i-1) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} 0 \\ a_{i-1}^{(i-1)} \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \gamma^{i-1} \\ 0 \\ \vdots \\ E_{i-1} \end{bmatrix}$$

$$R[i] \left\{ \begin{bmatrix} 1 \\ a_1^{(i-1)} \\ \vdots \\ a_{i-1}^{(i-1)} \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ a_{i-1}^{(i-1)} \\ \vdots \\ a_1^{(i-1)} \\ 1 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} E_{i-1} \\ 0 \\ \vdots \\ \gamma^{i-1} \end{bmatrix} - k_i \begin{bmatrix} \gamma^{i-1} \\ 0 \\ \vdots \\ E_{i-1} \end{bmatrix} \right\}$$

$$E_i = E_{i-1} - k_i \gamma^{i-1}$$

$$a_1^i = a_1^{i-1} - k_i a_{i-1}^{(i-1)}$$

$$\vdots$$

$$a_{i-1}^{(i)} = a_{i-1}^{(i-1)} - k_i a_1^{(i-1)}$$

$$a_i^{(i)} = 0 - k_i$$

Superscript = iteration

Also $\gamma^{i-1} - k_i E_{i-1} = 0 \quad \Rightarrow \quad k_i = \dfrac{\gamma^{i-1}}{E_{i-1}}$

$$k_i = \frac{R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(j)}{E_{i-1}}$$

LD Algorithm

$a_k^{(i)} \rightarrow k^{th}$ LPC coefficient value at i^{th} iteration

$E^{(i)} \rightarrow$ Residual error after i^{th} iteration.

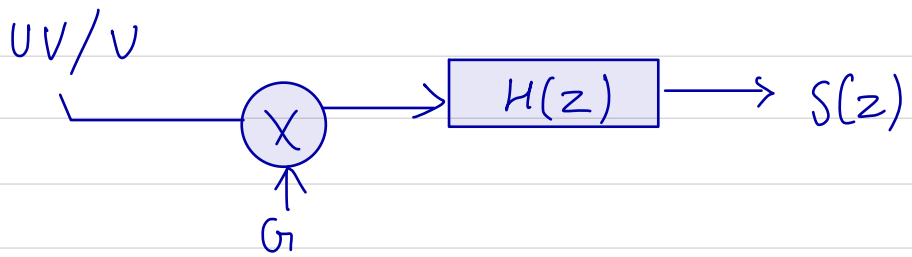
Steps: 1> Initially set $E^{(0)} = R(0)$

2> Calculate $k_i = \frac{[R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)]}{E^{(i-1)}}$

3> Set $a_i^{(i)} = k_i$ & $a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}$; $1 \le j \le i$

4> Calculate $E^{(i)} = (1 - k_i^2) E^{(i-1)}$

5> Repeat steps ②, ③, ④, until $i = p$ where p is linear prediction order.



$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}}$$

$a_k \hookrightarrow$ LP coefficient

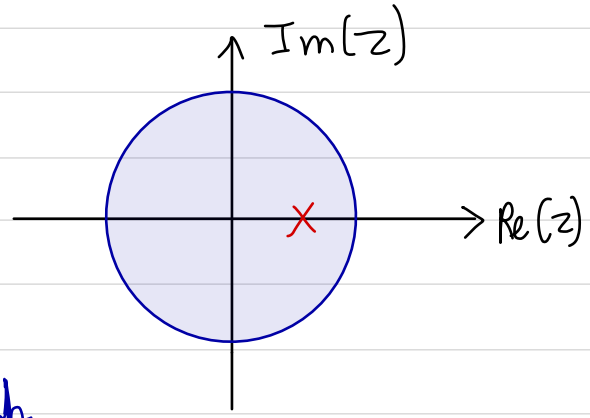
Freq domain : $z = e^{jw}$

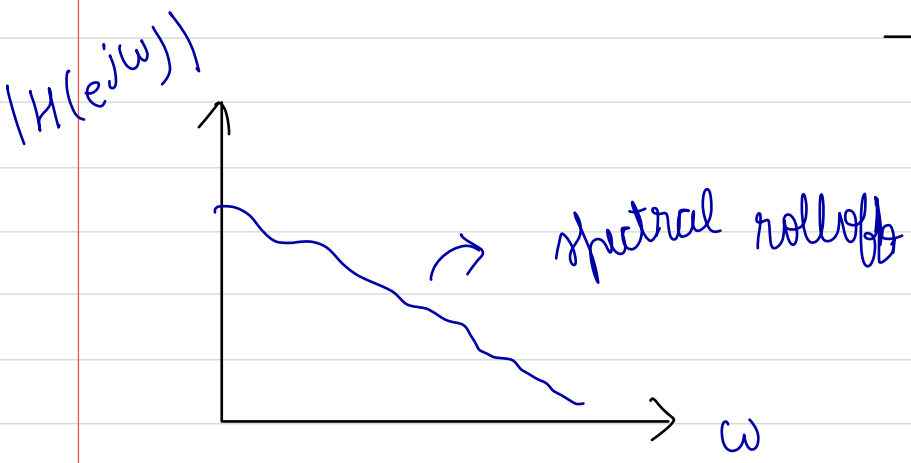
$$|H(e^{j\omega})| = \left| \frac{1}{1 + \sum_{k=1}^{p} a_k e^{-j\omega k}} \right|$$

Denominator : $1 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_p z^{-p}$

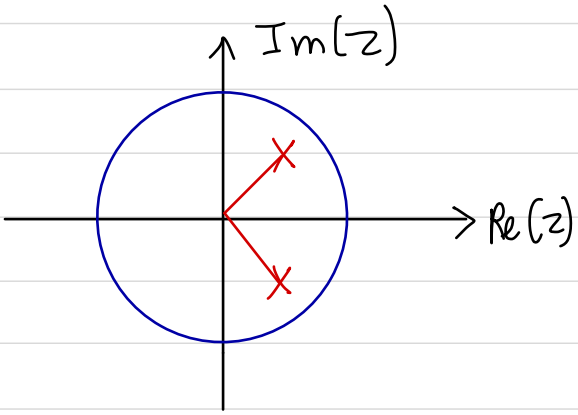
1) Let $p = 1$

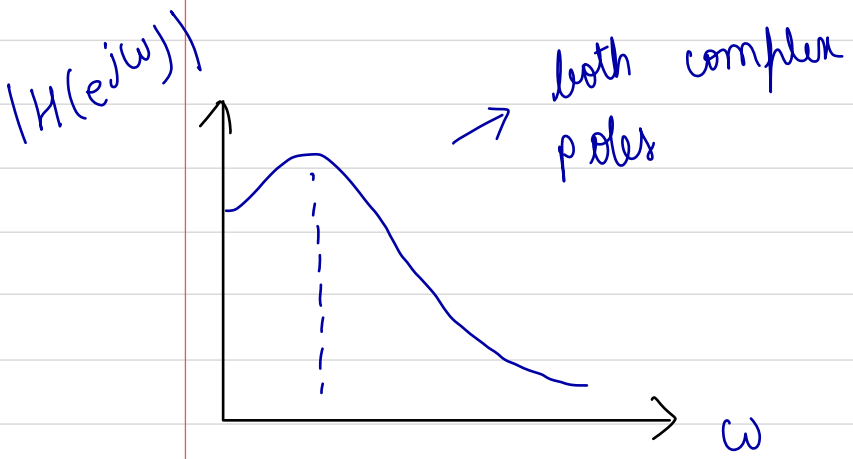
$$H(z) = \left| \frac{1}{1 + a_1 z^{-1}} \right|$$

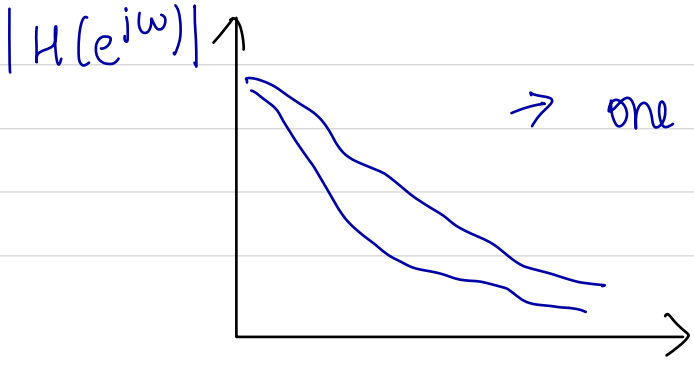




2) Let $p = 2$ → real & complex

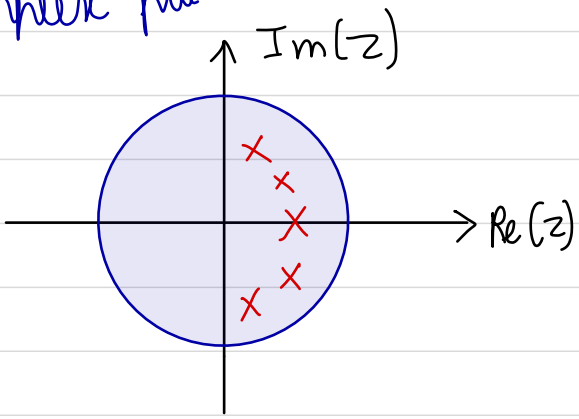


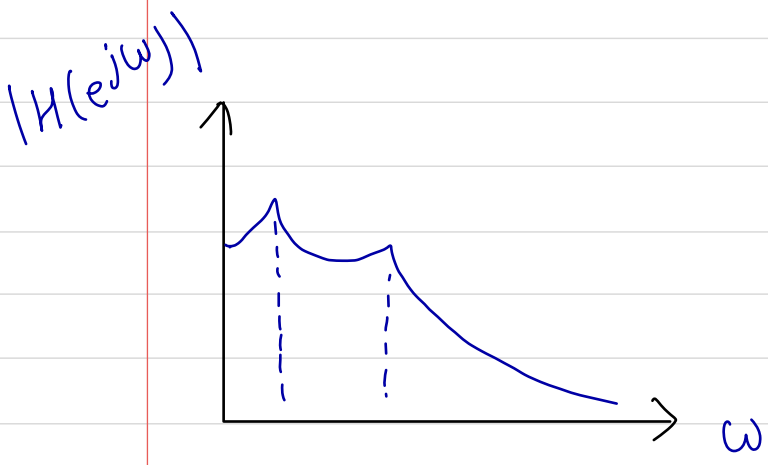




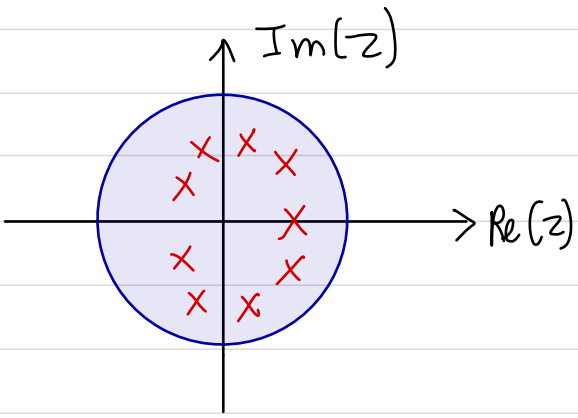
→ one real, one complex

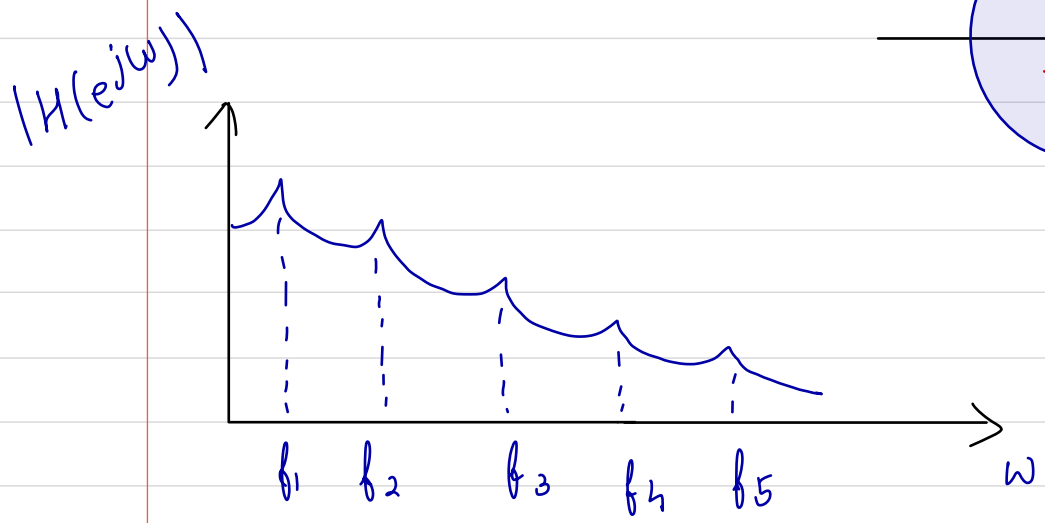
3> let $p = 5 \rightarrow$ 2 complex pairs

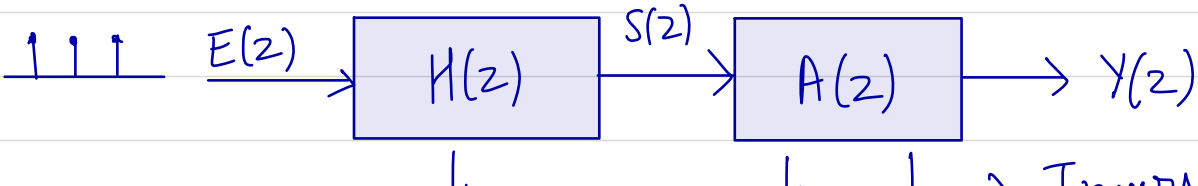




4> let $p = 11 \rightarrow$ 5 complex pairs

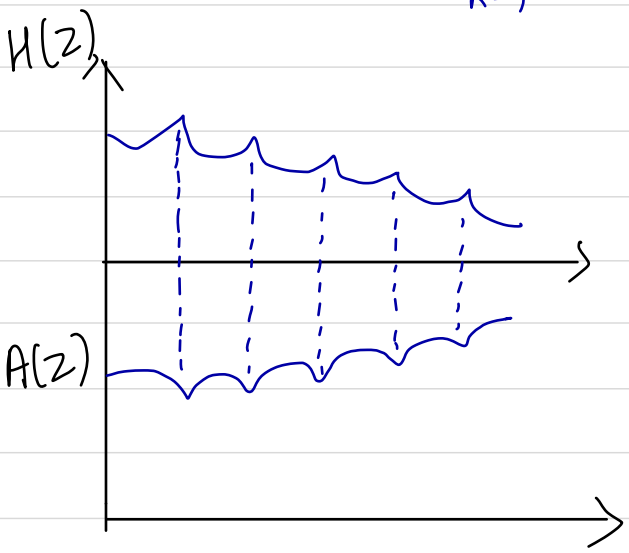






$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}}$$

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$$



LP residual

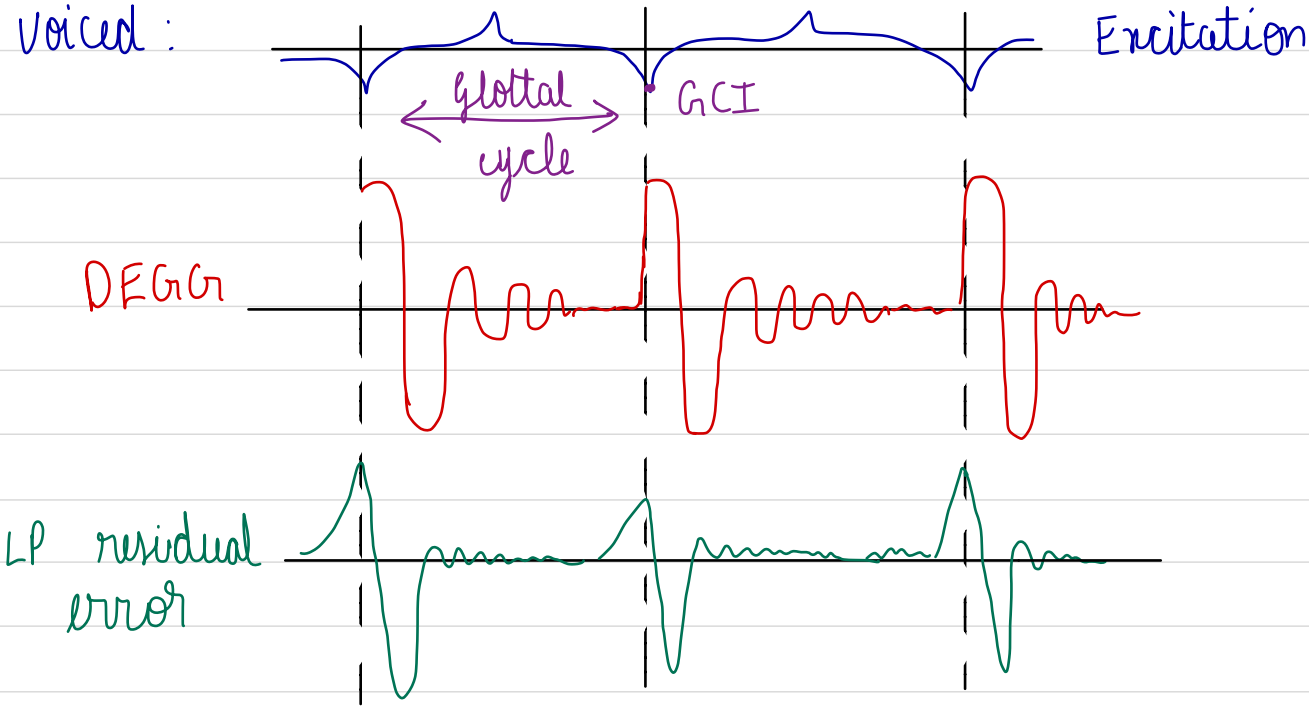
$$A(z) = \frac{Y(z)}{S(z)}$$

$$Y(z) = \left[1 + \sum_{k=1}^{p} a_k z^{-k}\right] s(z)$$

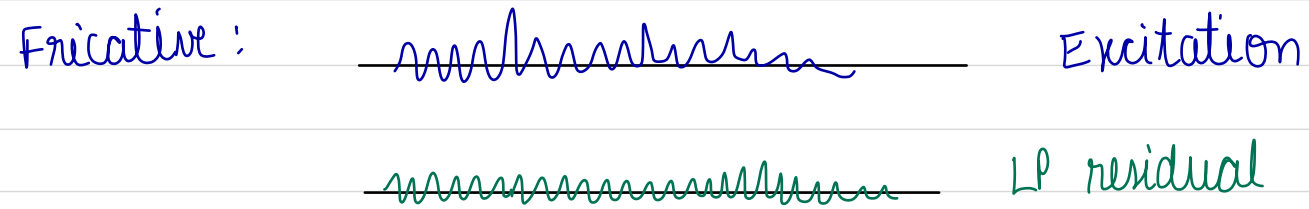
$$\text{MMSE} = s(n) - \tilde{s}(n)$$

$$= s(n) + \sum_{k=1}^{p} a_k s(n-k)$$

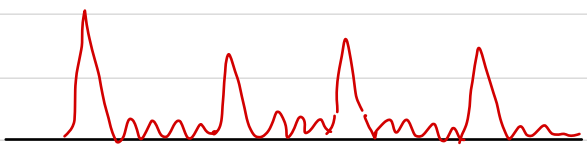
$$y(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k)$$



DEGGr = Differenced electrogastrogram



$s(n) \rightarrow (N \times 1)$



Assignment: 1> Compute LPC's ⇒ LD method
i/p : speech, LP order (p)
o/p : LPC (2D matrix

2> Plot LP spectrum
3> Compute LP residual & find out F_0

Gaussian mixture models

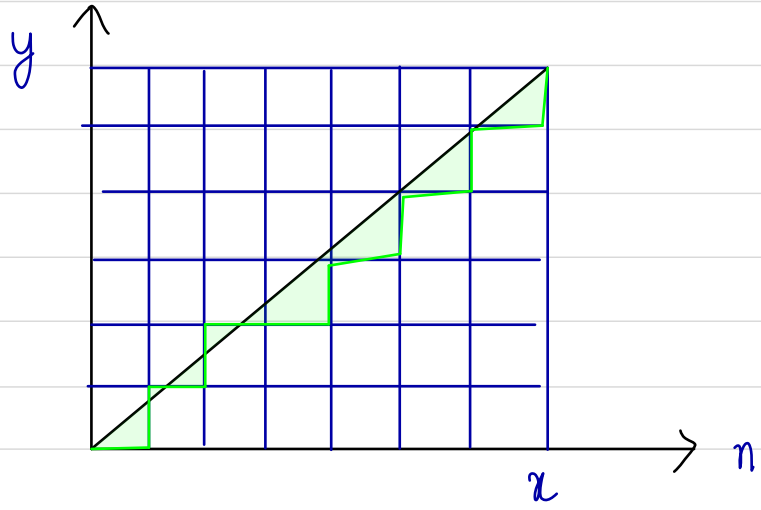
→ It estimates the data distribution into multiple gaussian models.
→ It is a generative model.

Different types of models:

1> K-Means
2> GMM - Gaussian mixture models
3> VQ - Vector Quantisation

Use cases:
1> Vowel v/s consonant classification
2> Keyword spotting

Dynamic time warping:



$W_1 \rightarrow 1, 2, \ldots T_x = 90$ features
$W_2 \rightarrow 1, 2, \ldots T_y = 130$ features

Need to map them properly. {since they have different number of features}

$\phi = \{\phi_x(k), \phi_y(k)\}$
↳ Non linear warping function

$$\phi(x, y) = \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)) \cdot \frac{m(k)}{m(\phi)}$$

$m(k)$ = path weighting coefficients
$m(\phi)$ = path normalization coefficients

$$d(x, y) = \min_{\phi} \{ d(\phi(x, y)) \}$$

Asynchronous Moves approach:

Bellmann optimality;

$\phi(i,j)$ – cost involved from city i to city j

$\phi(i,l)$ – cost involved from city i to intermediate city l

$\mathcal{E}_f(i,j)$ – cost of direct travel from i to j

$$\phi(i,j) = \min_{l} \left[\phi(i,l) + \mathcal{E}_f(l,j) \right]$$

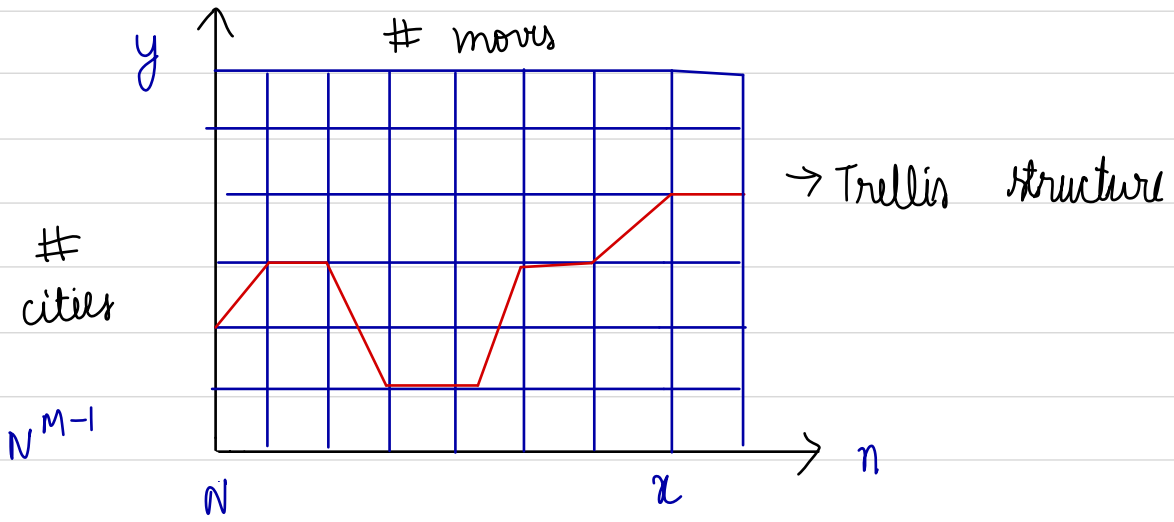
For direct travel (No halt) $\phi_1(i,j) = \mathcal{E}_f(i,j)$ where $i, j = 1, 2, \ldots N$

For one halt:

$$\phi_2(i,j) = \min_{1 \le l \le N} \left[\phi_1(i,l) + \mathcal{E}_f(l,j) \right]$$

$$\phi_3(i,j) = \min_{1 \le l \le N} \left[\phi_{3-1}(i,l) + \mathcal{E}_f(l,j) \right]$$

Synchronous mover approach



Viterbi Algorithm:

Recursive initialisation:

$$\phi(i, n) = \xi(i, n)$$

For path tracing: $\xi'(n) = i \quad , n = 1, 2, \cdots, N$

Recursive: $$\phi_{m+1}(i, n) = \min_{1 \le l \le N} \left[\phi_m(i, l) + \xi_p(l, n) \right]$$

$1 \le m \le N-1$

$$\xi'(n) = \arg \left[\phi_{m+1}(i, n) \right]$$

To do path tracing from $\xi'_{m+1}(n)$:

Constraints of template matching for Viterbi

1> Endpoint constraint:

- [sa] [re] [ga] [ma] -
- [ma] [ga] [re] [sa] -

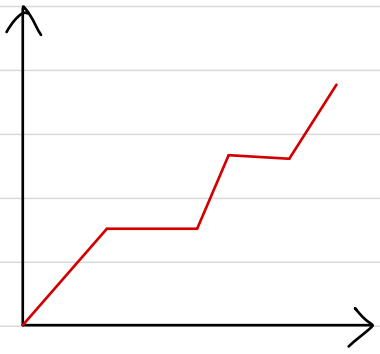
In endpoint constraint : $\phi_x(1) = 1 = \phi_y(1)$ for starting point

For endpoint:

$$\phi_x(T) = T_x$$
$$\phi_y(T) = T_y$$

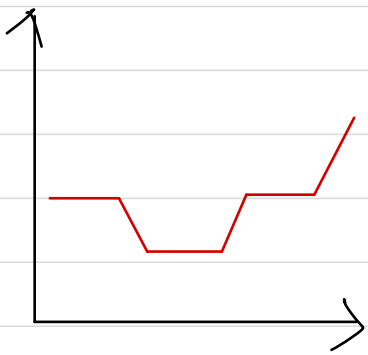
where T_x = number of time frames in X-direction
T_y = number of time frames in Y-direction

2> Monotonicity constraint



✓

monotonic



✗

non monotonic

So $\phi_x(k+1) \geq \phi_x(k)$
$\phi_y(k+1) \geq \phi_y(k)$

3> Local continuity constraint: (LCC)

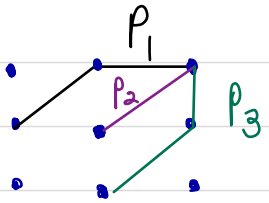
$$\phi_x(k+1) - \phi_x(k) \leq 1$$
$$\phi_y(k+1) - \phi_y(k) \leq 1$$

$$d_\phi(x,y) = \sum_{k=1}^{T} d\left(\phi_x(k), \phi_y(k)\right) \frac{m(k)}{M_\phi}$$

Normalization

Path weighing constant

Type a: $m(k) = \min\left[\phi_x(k) - \phi_x(k-1),\ \phi_y(k) - \phi_y(k-1)\right]$



P_1 : 2+1 = 3
P_2 : 1+0 = 1
P_3 : 1+2 = 3

P_2 is chosen as it is minimum.

Path normalizing factor :

$$m_k = \phi_x(k) - \phi_x(k-1)$$

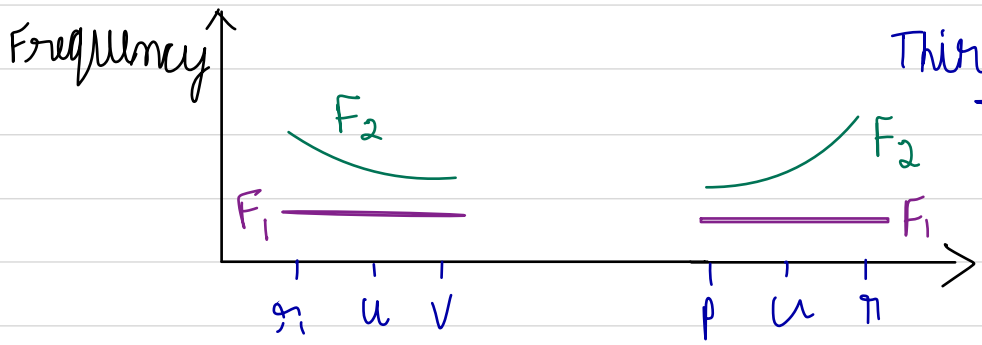
$$M_\phi = \sum_{k=1}^{T} m(k)$$

Hidden Markov models

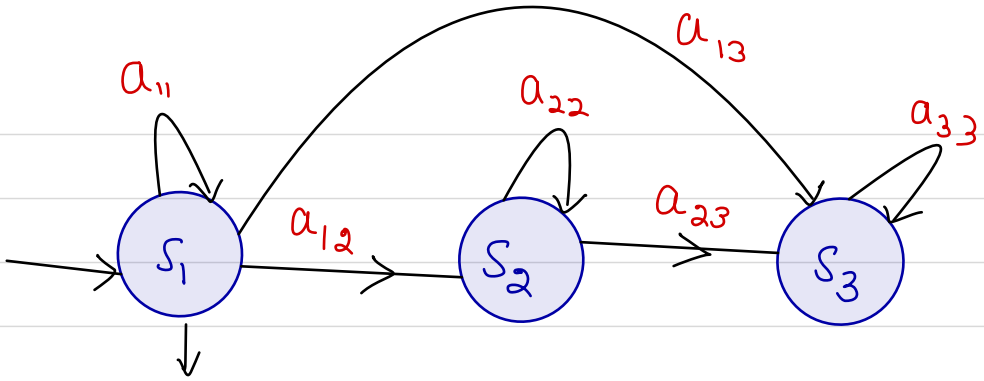
- In DTW N = large
- In HMM M <<< N

- In DTW parameter is mean
- In HMM parameters are mean & covariance

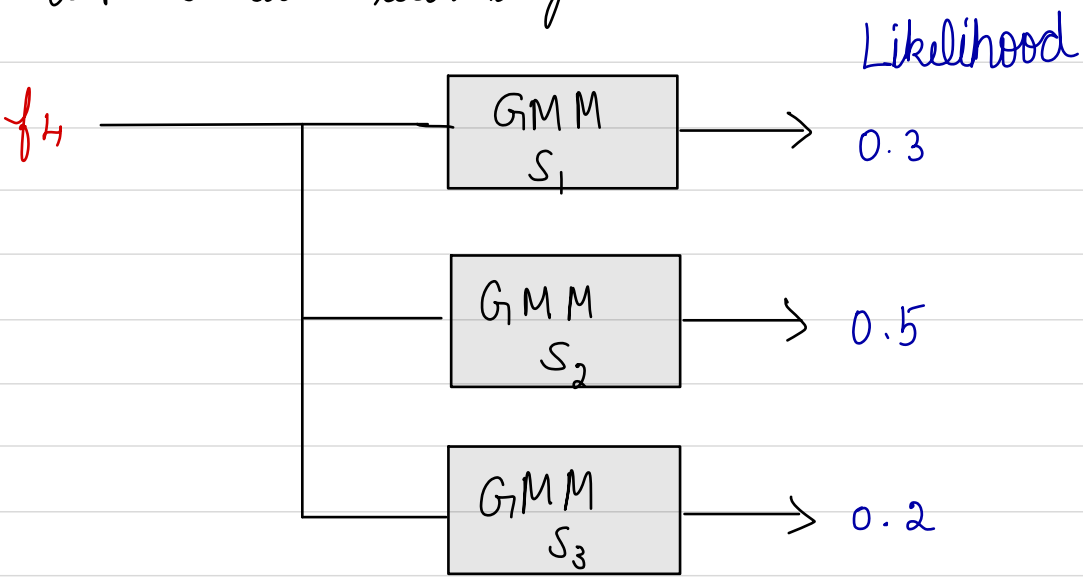
- Goal of DTW is to minimize distance.
- Goal of HMM is to minimize likelihood as it is a probabilistic model.



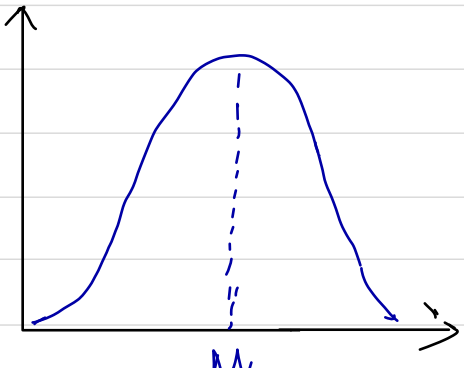
Thiruvanathapuram



State distribution learning



Gaussian mixture models



1D gaussian

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

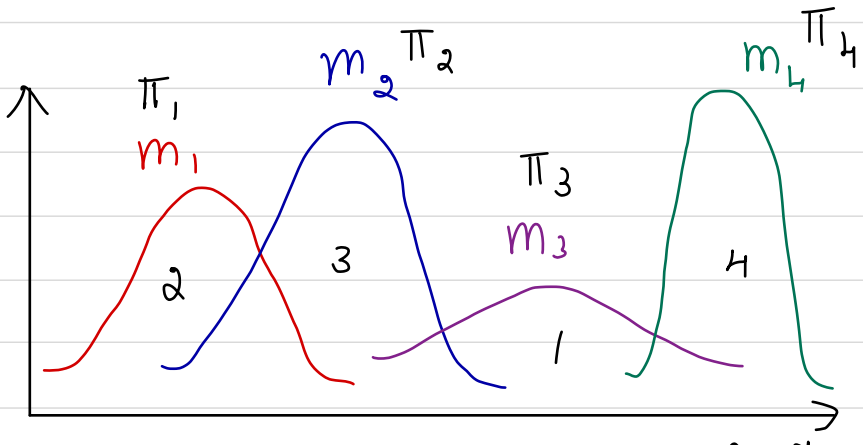
Multivariate gaussian distribution:

Eg: 39 D MFCC:

$$P(X|\vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left\{\frac{-1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)\right\}$$

$$\vec{\mu} = \begin{bmatrix} \mu_0 \\ \mu_1 \\ \vdots \\ \mu_{38} \end{bmatrix}_{39 \times 1}$$

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1N}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2N}^2 \\ \vdots & \vdots & & \\ \sigma_{N1}^2 & \sigma_{N2}^2 & \cdots & \sigma_{NN}^2 \end{bmatrix}$$



$\pi_1 = 0.2$ $\pi_2 = 0.3$

π_i = weights

$$P(X|\lambda) = \sum_{i=1}^{M} \pi_i P_i(x|\mu_i, \Sigma_i)$$

→ Probability distribution of i^{th} mixture

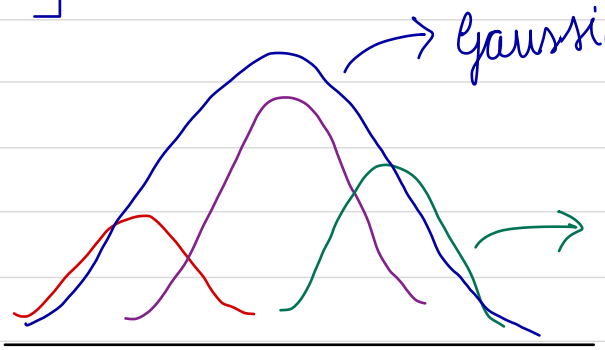
↓ weight of i^{th} component or mixture

Expectation maximization (EM) Algorithm :

→ Maximize the probability of generating feature vector from the model.

$$X = \begin{bmatrix} x_1 \end{bmatrix} \begin{bmatrix} x_2 \end{bmatrix} \begin{bmatrix} x_T \end{bmatrix}$$

$$E(X) = \begin{bmatrix} \mu_1 \\ \vdots \end{bmatrix} \mu$$



Procedure:

$X = \{x_1, x_2, \ldots\ldots, x_T\}$ let $T = 1000$

(39×1000)

$P(X|\bar{\lambda}) \geq P(X|\lambda)$

↓ model in the current situation

↘ model in previous iteration

$$p(i \mid x_t, \lambda) = \frac{w_i \, b(x_t \mid \mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k \, b(x_t \mid \mu_k, \Sigma_k)}$$

$i = 1, \ldots, M$

i^{th} Gaussian mixture

Current feature vector

Model GMM

Also $\sum_{i=1}^{M} p(i \mid x_t, \lambda) = 1$

Procedure:

$$\bar{w}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid x_t, \lambda)$$

$$\bar{\mu}_i = E(x) = \frac{\sum_{t=1}^{T} x_t \, p(i \mid x_t, \lambda)}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)}$$

w_i = weights

$$\sigma_i^2 = E(x^2) - (E(x))^2$$

$$= \frac{\sum_{t=1}^{T} x_t^2 \, p(i \mid x_t, \lambda)}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)} - (\bar{\mu}_i)^2$$

$$\bar{\Sigma}_i = \begin{bmatrix} \sigma_{11}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{22}^2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \sigma_{NN}^2 \end{bmatrix}$$

$\lambda = \{w_i, \mu_i, \Sigma_i\}$

$\bar{\lambda} = \{\bar{w}_i, \bar{\mu}_i, \bar{\Sigma}_i\}$

$p(X|\bar{\lambda}) - p(X|\lambda) \leq \text{Threshold}$ (eg : 0.0001)

↓

Termination condition

Digit recognition task:

Vocabulary : [0, 1, ..., 9]

↓

500 examples / class

→ Speaker variability / gender
→ Speaking mode
→ channel variability

Training : 500 x 10 = 5000 examples

MFCC : [39 x 1]

$X = \{x_1, x_2 \ldots, x_T\}$ let FV = 7000

↓ ↓ ↓

C_0 C_0 C_9

[39 x 7000]

GMM : $\lambda_0, \lambda_1, \lambda_2, \ldots \lambda_9$

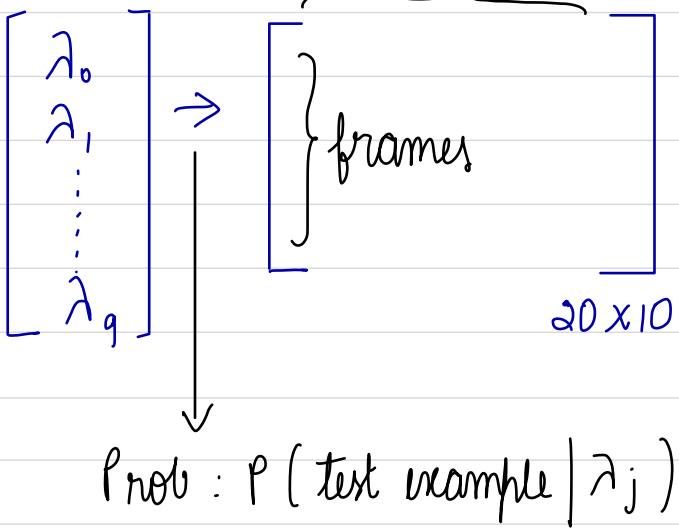
↓ ↓

(μ_0, Σ_0, w_0) (μ_1, Σ_1, w_1)

$fV \in C_0 \Rightarrow$ Kmeans $\rightarrow \mu_0^i, \Sigma_0^i, w_0^i \Rightarrow$ EM $\Rightarrow \mu_0, \Sigma_0, w_0$

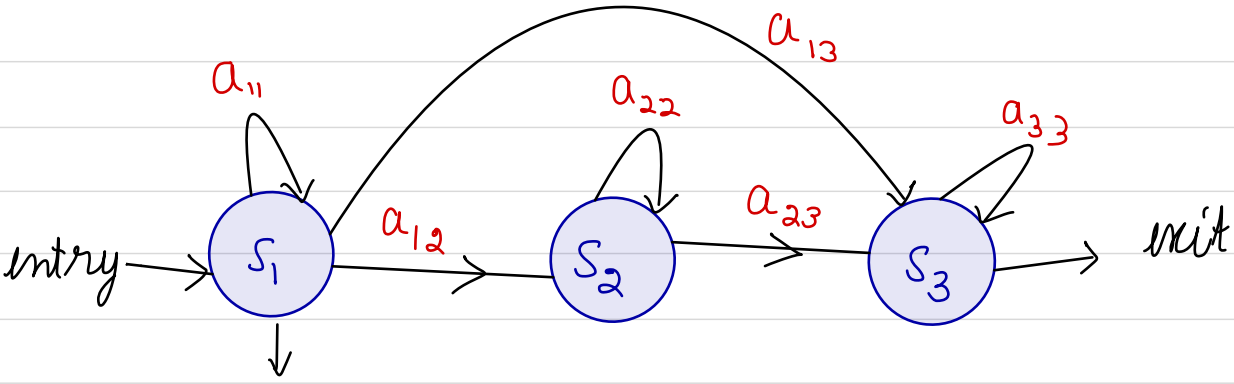
Testing → 100 X 10 = 1000 examples

Test example [39 x 20] → frames



Prob : P (test example | λ_j)

Transition probability



State distribution learning

$$P (q_t = j \mid q_{t-1} , q_{t-2} \dots)$$

First order hidden markov model

$P(q_t = j \mid q_{t-1} = i)$ is the objective.

Observable markov model (OMM):

Here we can see the past sequence clearly, like in the case of tossing a coin continuously.

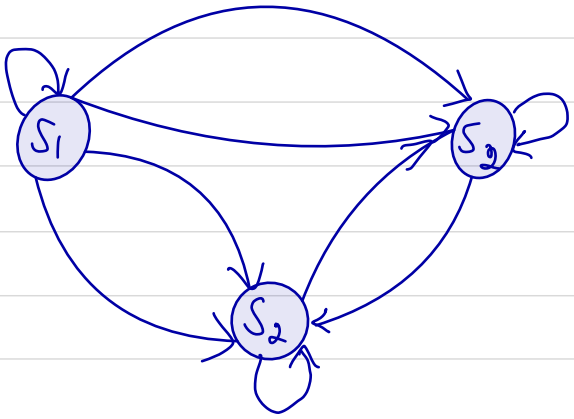
Eg: Take a biased coin with $P(H) = 0.7$ & $P(T) = 0.3$:

Observation: T, T, H, H, H, T, H, ... T
(T: 1, 1000)

Now consider 3 biased coins:

C_0	C_1	C_2
$P(H) = 0.3$	$P(H) = 0.2$	$P(H) = 0.6$
$P(T) = 0.7$	$P(T) = 0.8$	$P(T) = 0.4$

Observation: T, T, H, T .. T
Hidden: $C_2, C_1, C_1, C_2 \ldots C_0$



1> a_{ij} = state transition probability

→ final

→ initial

$$P(q_{t+1} = j \mid q_t = i)$$

2> N & M

N = states

M = observation

3> State Transition Matrix:

$$A = [a_{ij}]_{N \times N}$$

$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2N} \\ \vdots & \vdots & & & \\ a_{N1} & a_{N2} & \cdots & \cdots & a_{NN} \end{bmatrix}$$

$$\sum_{j=1}^{N} a_{ij} = 1$$

(row sum = 1)

4> States: modelled as — Histogram

— GMM

M distant Observation < H / T

$V = \{ v_1, v_2, \ldots, v_M \}$ ⇒ observation symbol

observation seq $O = \{O_1, O_2, \ldots, O_T\}$

$T = 1000$

Observation symbol probability:

$$b_j(k) = P(O_t = v_k \mid q_t = j)$$

probability of producing observation v_k at time t given that jth state at t.

$$\sum_{k=1}^{M} b_j(k) = 1$$

$$B = [b_j(k)]_{N \times M} \qquad j: 1 \to N, \quad k: 1 \to M$$

$$\boxed{\sum_{k=1}^{M} b_j(k) = 1} \rightarrow \text{stochastic constraint}$$

Initial State probability:

$$\pi_i = P(q_2 = i) \qquad i: 1 \to N$$

$$\sum_{i=1}^{N} \pi_i = 1$$

$$\lambda = \{A, B, \pi\} \quad (N, M) \quad \text{(parameters)}$$

Testing Procedure
Optimal State seq finding
Training Procedure

Problem 1:

Testing Procedure: The process of efficiently computing $p(O|\lambda)$, given the observation seq 'O' & model 'λ'.

Problem 2:

Optimal state seq finding: This involves finding the optimal state seq, given observation 'O' & model λ

Problem 3: Training Procedure:

Given observation seq 'O' & initial model 'λ' How to re-estimate model parameters.
$p(O|\bar{\lambda}) \geq p(O|\lambda)$

Consider "Mera Bhaarat Mahaan"

silence- (m) e r a bh aa r a t m a h aa n -sil

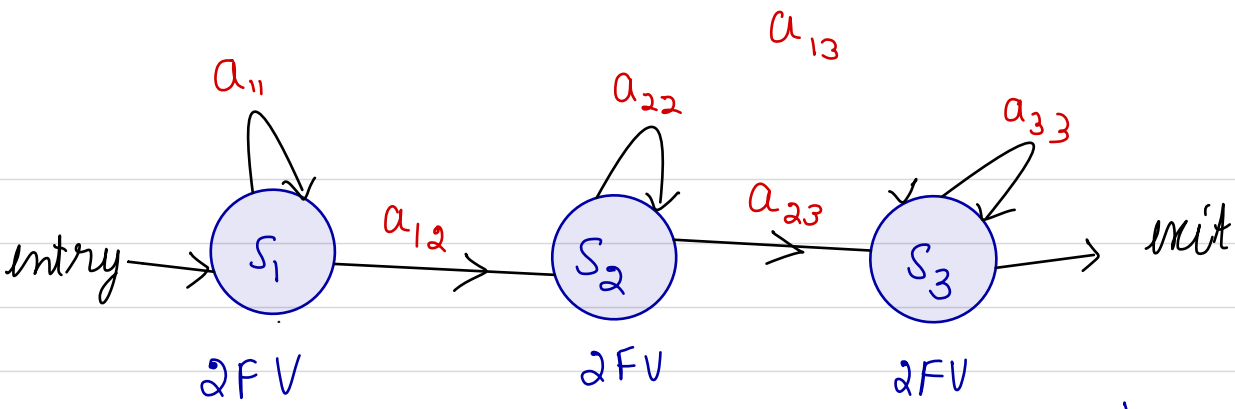
↓

3 state HMM

let 1 sec = 98 Frames (25 msec, 10 msec)

⇒ This will be [39 x 98] → 6 Feature vector/Phoneme

→ 2(FV / state) / Phoneme



$$a_{ij} = \begin{cases} 0.5 & \text{if } j = i \text{ or } j = i+1 \\ 0 & \text{otherwise} \end{cases}$$

$O = \{O_1, \ldots O_T\}$

$q = \{q_1, \ldots q_T\}$

$P(O|\lambda)$ → observation.

Solution to problem: Efficiently compute $P(O|\lambda)$

Direct Method / Brute force:

$P(O|\lambda)$ → Probability of 'O' being generated by λ.

$= \{O \ldots O_T\}$

$q = \{q_1, \ldots q_T\}$

$$P(O|q^t, \lambda) = \prod_{t=1}^{T} P(O|q^t, \lambda)$$

$$= b_{q_1}(O_1) \cdot b_{q_2}(O_2) \ldots b_{q_T}(O_t)$$

$$= \prod_{t=1}^{T} b_{q_t}(O_t)$$

$$P(O|\lambda) = \sum_{\text{all } q \to N^T} P(O|q,\lambda) \cdot P(q|\lambda)$$

$$P(q|\lambda) = P(q_1\ q_2 \ldots q_t|\lambda)$$

$$= [\pi_{q_1}\ a_{q_1 q_2} \ldots a_{q_{t-1} q_t} \ldots a_{q_{T-1} q_T}]$$

(q_1, q_2: S_1, S_2)

$$P(O|\lambda) = \sum_{\forall q \to N^T} \pi_{q_1} b_{q_1}(O_1)\ a_{q_1 q_2} \cdot b_{q_2}(O_2) \ldots a_{q_{T-1} q_T}\ b_{q_T}(O_T)$$

Forward procedure for $P(O|\lambda)$

Forward variable:

$$\alpha_t(i) = P(O_1, O_2, \ldots O_t, q_t = i|\lambda)$$

Initialization at $t=1$

$$\alpha_1(i) = P(O, q = i|\lambda) = \pi_i b_i(O_i)$$

↳ joint Probability

Recursion:

$$\alpha_t(i) = P(O_1, O_2, \ldots O_t, q_t = i|\lambda)$$

$$\alpha_{t+1}(j) = P(O_1, O_2, \ldots O_t, O_{t+1}, q_{t+1} = j|\lambda)$$

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i)\ a_{ij}\ b_j(O_{t+1})$$

$1 \le t \le T-1$

$1 \le j \le N$

Termination:

$$\alpha_T(j) = \sum_{i=1}^{N} \alpha_{T-1}(i)\, a_{ij}\, b_j(O_T)$$

$$p(O|\lambda) = \sum_{j=1}^{N} \alpha_T(j) \qquad p(O_1, \dots O_T, q|\lambda)$$

Method 3: Backward procedure:

Backward variable:

$$\beta_t(i) = P(O_{t+1}, O_{t+2} \dots O_T | q_t = i, \lambda)$$

Initialization: $\beta_T(i) = P(O_T | q_t = i, \lambda)$

Recursion: $\beta_t(i)$

$$\beta_{t+1}(j) = P(O_{t+2} \dots O_T | q_{t+1} = j, \lambda)$$

$$\beta_{t+1}(j) = \sum_{j=1}^{N} a_{ij}\, \beta_t(j)\, b_j(O_{t+1})$$

$1 \le t \le T-1$

$1 \le i \le N$

Termination: $\beta_1(i) = \sum_{j=1}^{N} a_{ij}\, \beta_2(j)\, b_j(O_2)$

$$P(O|\lambda) = \sum_{i=1}^{N} \beta_1(i)\, \pi_i\, b_i(O_1)$$

Problem 2: Optimal state sequence estimation

given $O, \lambda \rightarrow$ find out q

$$\delta_t(i) = \max_{q_1, q_2 \dots q_t} (q_1, q_2, \dots, q_t = i, O_1, O_2 \dots O_t | \lambda)$$

Path tracing : $\psi_t(j) = \arg \delta_{t-1}(j)$

Initialization :

$$\delta_1(i) = p(q_1 = i, O|\lambda) = \pi_i b_i(O_1)$$
$$\psi_1(i) = 0$$

Recursion :

$$\delta_{t-1}(i) = \max p(q_1 q_2 \dots q_{t-1} = i, O_1, O_2 \dots O_{t-1}|\lambda)$$

$$\delta_t(j) = \max p(q_1 q_2 \dots q_t = j, O_1 \dots O_t | \lambda)$$

$$\delta_t(j) = \max_t \delta_{t-1}(i)\, a_{ij}\, b_j(O_t)$$

$$\psi_t(j) = \arg \delta_t(i)$$

Termination: $\delta_T(j) = \max \delta_{T-1}(i)\, a_{ij}\, b_j(O_T)$

$$\psi_T(j) = \arg \delta_T(j)$$

$$i = \max_j \delta_T(j) \qquad q^* = \psi_{t+1}(q^*_{t+1})$$
$$1 \le j \le N$$

Problem 3: Training of HMM

$$\lambda = \{A, B, \pi\}$$

$$\gamma_t(i) = p(q_t = i|\lambda)$$
$$= p(q_t = i|\lambda) - \frac{p(O|\lambda)}{p(O|\lambda)}$$
$$= \frac{p(q_t = i, O|\lambda)}{\sum_{j=i}^{N} p(q_t = j, O|\lambda)}$$

$$\gamma_t(i) = \frac{P(O_1 \dots O_t, q_t = i \mid \lambda)\, P(O_{t+1} \dots O_T, q_t = i \mid \lambda)}{\sum_{i=1}^{N} P(q_t = j, O \mid \lambda)}$$

$$\gamma_t(i) = \frac{\alpha_t(i)\, \beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\, \beta_t(j)}$$

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j \mid \lambda) \left(\frac{P(O \mid \lambda)}{P(O \mid \lambda)} \right)$$

Vowel identification > HMM
Speaker recognition

$$\xi_t(i,j) = \frac{P(q_t = i, q_{t+1} = j, O \mid \lambda)}{\sum_{i=1}^{N} \sum_{j=1}^{N} P(q_t = i, q_{t+1} = j, O \mid \lambda)}$$

$$= \frac{\alpha_t(i)\, a_{ij}\, b_j(O_{t+1})\, \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i)\, a_{ij}\, b_j(O_{t+1})\, \beta_{t+1}(j)}$$

$$\bar{\pi}_i = P(q_1 = i) = \gamma_1(i)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$P(O \mid \bar{\lambda}) \geq P(O \mid \lambda)$$

$$\bar{b}_j(k) = \frac{\sum_{t=1}^{T} \gamma_t(j)\, \delta(O_t, v_t)}{\sum_{t=1}^{T} \gamma_t(j)}$$