② Zero crossing Rate: (ZCR)

Unvoiced has high

Voiced has low

③ Presence of pitch:

Voiced has high pitch

Unvoiced has low pitch. (Extraction).

④ Auto-correlation.

Voiced has high auto-correlation coefficient

Unvoiced has low auto-correlation coefficient

⑤ Band Ratio : $\dfrac{\text{Energy upto 1KHz}}{\text{Total energy}}$

High for voiced

Low for unvoiced.

⑥ Linear prediction error :

$\hat{s}_n[k] = \Sigma a_k s[n-k]$

$\hat{s}[n] = \Sigma a_n s[n-k]$

$\boxed{\text{error} = s[n] - \Sigma a_k s[n-k]}$

more for unvoiced
less for voiced } → Explanation:
Autocorrelation
coefficient.

Normalized LP error analysis :

⑦ First LP coefficient:

Consider a_1,

$a_1 \uparrow$ in voiced than unvoiced because voiced has high autocorrelation coefficients than in unvoiced.

Pitch modification }
Time scale modification } Trade off

modify time-scale pitch changes
modify pitch, time scale changes.

Voiced and unvoiced speech detection:

three types of regions:

- Voiced — all vowels
- unvoiced
- Silence. — noise

Voiced region has better features.

Speech coding (AMR) — adaptive multi-rate coder.

64Kbps : 4.4Kbps - 122kbps
(Silence) (Voiced)

Pitch detection

Frontend block for speaker recognition.

Voiced vs unvoiced detection algorithms :

① Energy Threshold :

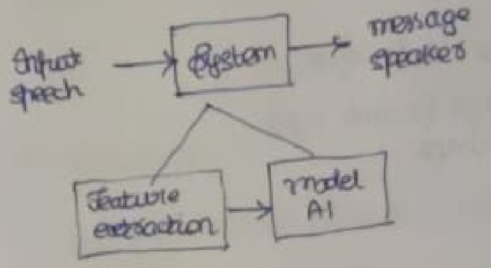
Voiced ↑
Unvoiced ↓

Drawback: Noise also have high energy and will be detected as unvoiced / voiced signal.

Not suitable method.

LP spectrum — Envelope of the FFT
↓
outputs envelope.

Speech Analysis :



Speech analysis:

Segmential analysis (10-30ms) }
Sub-Segmentical analysis } differentiated wrt length of window.
Supra-Segmential analysis }

<5-10ms

>50ms-100ms

Segmental is most popular because the duration can cover around 2 to 3 pitch cycles

20ms = Good frame size
2 to 3 pitch cycles.

Segmental analysis — VTS, Excitationer

SubSegmental analysis — Excitationer

Supra Segmental analysis — Prosoudy

Pitch will not be there when the speaker is unvoiced / not talking
Pitch can vary within a person.

Short time Fourier transform works good if signal in my window is stationary.

Frame size??

FT works if whole signal is stationary. But speech is dynamic.

Spectrogram — Implements STFT

STFT: for a signal STFT is not unique, depends on window size (frame size). window shape

Usually, good frame size ~20ms.

LTI model of speech:

Linear prediction:

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

usually $p = 8$ or 10.

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

actual − predicted.

Z transform

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

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

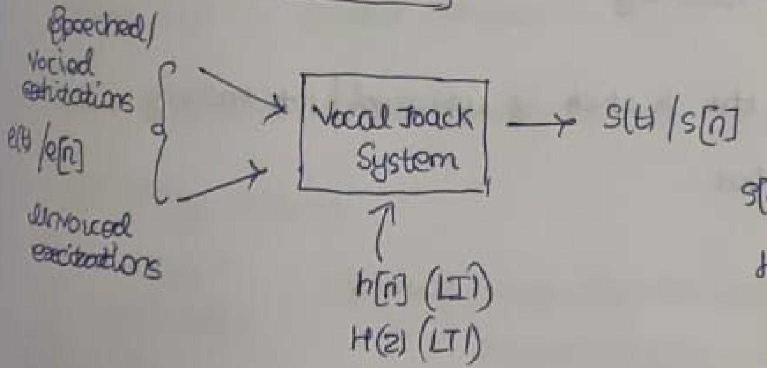
$$\frac{E(z)}{S[z]} = 1 - \sum_{k=1}^{p} a_k z^{-k}$$

a_k values can be obtained by minimizing error $e[n]$.

$$\hat{e}[n], \ \frac{d}{da_k}\left(\hat{e}^2[n]\right) = 0$$

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

→ Transfer function of vocal track system.



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

$$H(z) = \frac{S(z)}{E(z)}$$ Synthesis equation

$$\Rightarrow \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$ (all pole system)

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

(analysis equation)

(all zero system)

Semivocals: Vocal tract is not completely closed nor completely open.

య ర ల వ
y r l v

Fricatives : Friction in the excitation part

h sh shh h s

Co-articulation :

Before one sound production ends, other sound production starts

Consider (అడిగి, అతడు) —— Co-articulation.

Both are different

Formants :

air → vocal folds vibrations → → $\boxed{\text{VTA}}$ → Resonate frequency

These frequencies are more or less same for a phone by every person.

Frequencies of a syllable are approximately same produced by any person.
These are called formants.

Lecture : 4

Recap: Vocal tract system,
excitations
most important things of speech production.

Vowels = 150-300ms, Consonants duration = <120ms

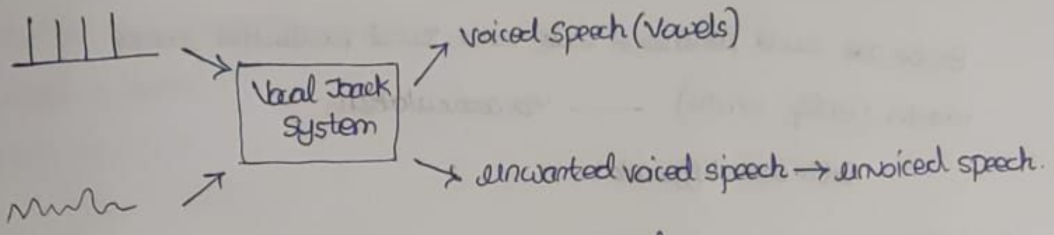
Consider a signal with vowels a,e,i
and another signal with vowels e,i,a
How to identify the vowels
Fourier transform ? Not useful because FT spectrums of both the signals are more or less the same

Answer: STFT
Short time Fourier transform.

manner of articulation —— excitation characteristics

Place of articulation —— system (vocal tract) characteristics.

Generally,
vowels have high strenght and long duration
consonants have low strength and low duration



For all vowels, vocal tract system must be opened.

All vowels are voiced and have no constriction.

For consonants:

POA: Place where you put a constriction in producing a consonant in vocal tract system.

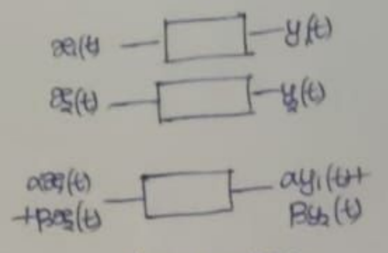
	manner of articulation					
Place of articulation	Unvoiced		Voiced		Nasals	Semivowel
	unaspirated	aspirated	unaspirated	aspirated		
Velar	k	kh	g	gh	kn	
Palatal	ch	chh	j	jh	chn	y
Alveolar	T	Th	D	Dh	Tn	r
Dental	t	th	d	dh	n	l
Bilabial	p	ph	b	bh	m	v

* Phone is the basic sound production unit

Syllable is the combination of phones.

90% Indian syllables are consonant vowel combination. (ka, tha, pa)

Linear and Time invariant systems:


$x_1(t)$
$y_1(t)$
$x_2(t)$
$y_2(t)$
$\alpha x_1(t) + \beta x_2(t)$
$\alpha y_1(t) + \beta y_2(t)$

Linear system

$x(t)$ — $y(t)$

$x(t-t_0)$ — $y(t-t_0)$

time invariant system.

Z-transform:

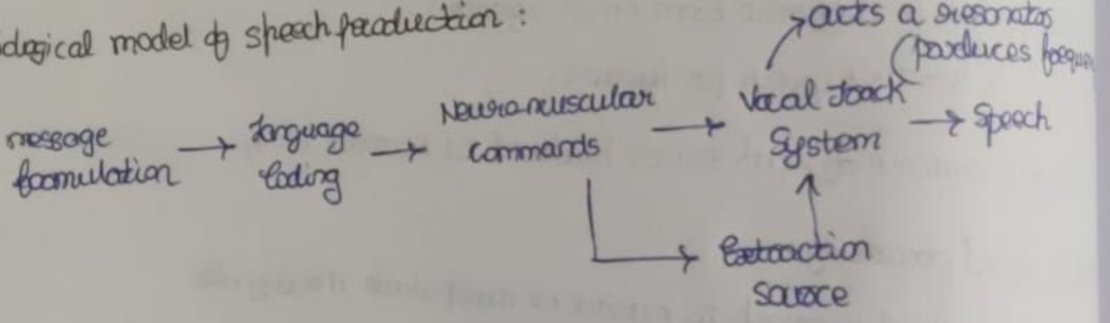
$X(z) = \Sigma x[n] z^{-n}$ $\boxed{z = re^{j\theta}}$

r is added because sometimes $\Sigma x[n]$ is not summable but $\Sigma x[n] r^{-n}$ may be summable.

When $r=1$, $z = e^{j\theta}$, $X(z)$ becomes discrete time fourier transform (DTFT).

assume if only $X(z)$ is given, it is insignificant without ROC.

Physiological model of speech production:

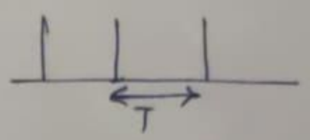


larynx: acts as exhitationer

Rate of vocal fold vibration is called speech (sorry pitch)

children > female > male

Vocal folds, when they are open, air passes through vocal folds as an impulse.



$\frac{1}{T}$ = pitch = rate of vocal fold vibrations.

Each impulse = epochs / ISE.

Speech Signal Processing.

Grading policy:

Quiz – 10%
midsem – 20%
assignments – 20%
Project – 25%. (10%-mid, 15%-end)
Endexam – 25%.

- speech basics
- → speech applications + project
- → speech analysis.

why speech processing needed?

Human Computter interaction.

Easy accesable of technology to all classes of people.

How speech is different from other signals?

↳ Involves speech production.

Legal sequence of legal sounds produced by human being.

what is signal processing:

mathematical approach to extract or manipulate the signals.

Fourier transform:

$$\left.\begin{aligned} X(\omega) &= \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\,dt \\ x(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{j\omega t}\,d\omega \end{aligned}\right\}$$

continuous fourier transform

0-4KHz sounds are understandable by humans.

$$x(t) \xrightarrow{\text{sampling}} x[n]$$

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

$$x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}\,d\omega.$$

Sampling DTFT = DFT

$$X(k) = \sum_{0}^{N-1} x[n]e^{\left(-j\frac{2\pi}{N}\right)kn} \qquad x[n] = \frac{1}{N}\sum X(k)e^{j\frac{2\pi}{N}kn}$$