Teacher's Signature

Lecture 01 (11 Aug).

Speech: legal sequence of legal sounds produced by humans. understandable by other humans

Broad goal of: speech processing
- Human–machine interaction as natural as possible

ASR → automatic Speech Recognition
- Speech → text

Applications of Speech processing

1. ASR
2. Language recognition
3. Speaker Recognition
4. Speech synthesis — Text to speech (TTS) eg. GPS
5. Speech analysis
   - Event detection
   - Speech to features | extract useful features
6. Speech Coding : since the wireless channel bandwidth is limited so we

Analog signal → Sampling → Quanti-sation → Digital signal

Sampling: how to sample the continuous signal into discrete signal

Speech → Speech coding → Compressed speech

2 FM (max frequency) — WHY? — to avoid aliasing (loss of info)

FM = 4 KHz

each sample represented using 8 bits

if 8 kHz samples × 8 bits/sample

⇒ bitrate 64 Kbits/sec

⇒ 64 Kbps bandwidth

but since speech has a lot of redundancy we can reduce the bitrate to (eg. 12.2 / 13.3 Kbps) GSM and then reconstruct back from 12.2 to 64

this is called speech coding

Teacher's Signature

Now, we can accomodate 5 users instead of 1 — because we reduce bitrate to 1/5th.

5G uses Enhanced Voice Serieves (EVS)
Upto 3G/4G — AMR Coder
Adapt to multiple codec
codes the speech signal depending on the quality of speech
If it is very imp. frame it uses high bitrate.

Bitrate — (5.7 to 12.2 kbps)

unwanted speech

→ Speech coding uses LP Analysis.

(7) Speech Enhancement
↳ reduce noise

3 kinds of noise
a. Bg noise
b. Reverberation
c. multispeaker noise

SHREE DATE: / / PAGE NO.:

$f_s = 30$ kHz

→ Speech recognition — HMM (baseline)

→ Speaker recognition — Variants of gaussian models (baseline)

Now, some NN is baseline
eg. ASR - RNN or Seq to Seq
Speaker recog. - DNN + attention / multiheaded attentn

Lecture 02 (14 Aug)

→ Signals - expressed as funcn of independent variable

Continuous signal	Discrete signals
→ Continuous values	→ Discrete values

Analog to Digital

① Sampling (Analog to discrete)

(a) Analog signal × uniform impulses → discrete signal

Teacher's Signature

→ Sampling rate matters when reconstructing the signal back

→ Sampling — highpass f_1 to ∞

/

lowpass sampling — 0 to max frequence (f_m)

bandpass sampling — f_1 to f_2

sampling rate = F_s

$$f_s \geq 2 f_m$$

→ time domain —FT→ frequency domain

→ (a.) spectrum in frequency domain X spectrum of impulse

Spectrum becomes repetitive with f_s

So if $f_s < 2f_m$ then aliasing happens and hence signal lose (*) PTO X2

and if $f_s \geq 2f_m$ even though they repeat, they won't touch each other

SHREE
DATE: / /
PAGE NO.:

Teacher's Signature

Nyquist Criteria

$$f_s \geq 2f_m$$

sampling frequency $\geq$ 2 (higher frequency content)

A continuous time signal can be completely represented in its samples and recovered back if the sampling frequency is 2 wice of highest frequency

In speech we take,
min. sampling rate = 8 kHz
because we assume our sounds have frequency upto zero to 4 kHz.

So max freq = 4 kHz (f_m)
So f_s as 8 kHz minimum (f_s)

Some sounds have frequencies upto 8 kHz so ideal f_s = 16 kHz but due to bandwidth limit 8 kHz is used.

PAGE NO.
DATE / /

Teacher's Signature

HD quality sound use high ~~bitrates~~ sampling rate i.e. more samples are used to represent the signal.

(2) Quantisation (discrete to digital signal)

no. of bits used to represent a sample

if more no. of bits is used - i.e. more bitrate - the quantisation error reduces

more Bitrate = high sampling rate and more no. of bits

information is preserved hence quality is good.

we need low quality also to fit to a given bandwidth.

LTI (Linear Time Invariant)

- Linear — homogenity & superposition
- Time-invariant

Homogenity

$a x_1 + b x_2 \Rightarrow a y_1 + b y_2$

Superposition

$x_1 \Rightarrow y_1 \qquad x_2 \rightarrow y_2$

$x_1 + x_2 \rightarrow y_2$

Impulse response — output of impulse

Output for LTI is convolution of input & impulse response.

(II) Discrete time → convolution sum
Continuous time → convolution integral

Fourier series — for periodic

Fourier Transform — non periodic & periodic

→ To know characteristics of system/signal we multiply it by an impulse response (train of impulses).

Teacher's Signature

Aliasing

when aliasing occurs — i.e. when there is an overlap in signal low frequencies will get affect which are the voiced regions.

In discrete domain

output = convulatn (input, system response)

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

Teacher's Signature

SSP

Class: 21st Aug : Lecture 04

	energy	duration	harmonic
→ voiced consonant	lower	low	uniformity
unvoiced consonant	lowest	low	random
vowels	highest	highest	pattern

→ Formants — in similar range for vowels

→ Spectrum section — FFT of the spectrum

→ Linear Prediction

→ LPC — smoothening of spectrum

→ in FFT — peak picking is difficult while it isn't so in LPC.

→ LPA fails in unvoiced consonants due to random (non-uniformity) & in aspirated & other sound but since speech is dominated by VC & vowels it is feasible

→ Vowels → Vocal fold vibration (

↓

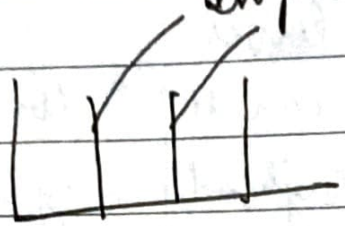
rate : Pitch

→ EGG Electroglottalgraph

extension in wavesurfer.

→ Impulse — open

Regular Intervals — open & close



→ Diff b/w impulses — pitch period

→ Impulse train ← assuming that voiced sounds are like impulse train

→ Glottal closure instant / Instant of significant excitation

→ In Diff EGG — the negative peaks are epoch locations

Epoch

→ LP — predict current sample because there is periodicity in picture

→ Frame size - 20 ms — 2 to 3 pitch cycles

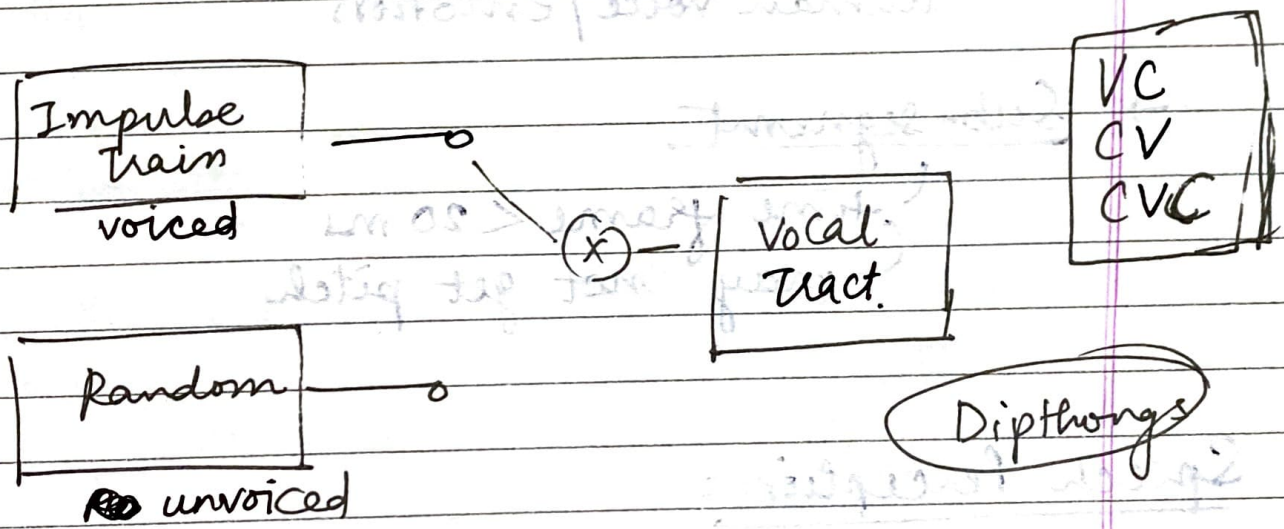
→ {P = 8 or 10} — in LP.

→ Z-transform — Study.

→ Residual — excitation

→ all pole filter because if we inverse the "forward filtering" equation — there are all poles.

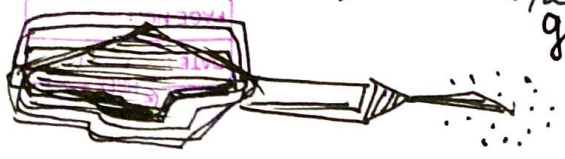
→



→ Plosive — MOA — if phone is blocking the air

→ Spectral details — f_0, f_1, f_2 } write formants

Formant tracking — yellow, red, green } mark range (Pg 13 Table 5)



Teacher's Signature

Lecture 05 (25th Aug)

Speech analysis

extract features to build speech systems

most speech sys use segmental

Q What features to use to automise phonetic transcription? And how to get these features?

Ans ~~a. Infoemat from frequency bands~~

a. frequency bands

b. ~~spectral~~ energy bands

~~b. Pitch Energy~~

Speech analysis techniques – depending on frame size

① Segmental analysis (around 20-50 ms)

use spectral features or frequency related features ← related to vocal tract

→ 20ms corresponds to 2-3 pitch cycle & signal won't change much. i.e. assume that stationary signal

② sub segmental (around 5ms)

frame size < 20 ms

use → extract excitatn source related features like epoch-related features (MOA)

may have redundancy (i.e. Excit source feature have redundancy comp. to vocal tract feature)

③ supra-segmental (imp to human but not to speech systems)

frame size > 100 ms

use → prosidy (long-term features) : extract energy patterns, durat patterns, pitch patterns } have natural variatn

Teacher's Signature

Class : 28 Aug (Lecture 06)

→ Segmental technique is popular (TF = 20 ms)
with spectral features

→ Supra segment
(prosodic features
(pitch ~~duration, energy~~ contour,
energy & durat^n of wave)
↓
related to human voice/emotion

→ Sub-segment
(time frame < 20 ms
(may not get pitch
excitat^n feature

Speech Perception :

- Voiced / Unvoiced detection

outer ear → middle ear → inner ear { neuro muscular commands
↓
converts to neuro commands

→ speech understanding also depends on language / context understanding

Voiced/Unvoiced detect^n — is used in most speech systems + first step - before extracting segment features

Applicat^{n}s : → speech coding
→ voiced activity detect^n (VAD) — activate Alexa
→ AMR
→ Emotion/ language / Speaker identificat^n

→ zero-crossing is less in voiced.
→ energy is more in voiced
→

~~Voiced activity~~

→
→ more the bit rate — better the coding
(AMR) → Adapt to multirate codec — supports multiple bit rates

4.75 to 12.2 — most popular
(lowest bit rate) (highest BR)

→ Frame size - 20ms.

→ if voiced - use highest bit rate
if silence - use lowest bit rate

→ AMR uses VAD

Q Why removing unvoiced/silence part & extract features only from voiced speech while performing speaker recog/emotn etc.?

→ But for speech recognition — silence and unvoiced also important

Ans. Only voiced regions have information about the excitation source / vocal tract characteristics.

Silence will be common across all speakers — it is redundant information.

In silence & unvoiced speech the vocal tract system is not active. And they may confuse the system very

Prosody info isn't useful for speech systems. Prosody may change for the same speaker.

Frame-level / Block processing of speech

Used mostly for speech processing

Teacher's Signature

Voiced/Unvoiced Detect^{n}.

Basic approaches & pros & cons and formula :

① Easiest : Energy (for one (100 ms) frame or 3-4 (multiple) frames

→ 20/50/100 ms
→ Compute energy or log energy
→ put threshold ← greater than this — voiced
0 — silence
less — unvoiced

Pros	Cons
→ easiest & simple	→ Noise ~~can~~ nt it ~~effect~~ will give wrong results
→ if clean speech, & if we put hard ↓ easy	→ we aren't exploiting any characteristics of voiced/unvoiced
	→ dynamic putting a threshold is difficult ∴ we cannot put a static threshold.
	→ some voiced consonants may get removed if not enough data is available

Teacher's Signature

(2) Zero crossing — no. of times signal crosses zero

↓
more
for unvoiced

Cons: if noise ⊕nt, it may look like noise & noise will also have more zero crossing

Pros: - putting up a threshold may be slightly easier but fails in case of noise
- easy & simple

→ Spectral energy is more for lower frequencies. for voiced & vice versa.

(3) Spectral energy
Lower Energy
Total band energy

more for voiced
less for unvoiced

→ more efficient till now
→ Spectral energy may sometimes have an issue of threshold

Teacher's Signature

④ Periodicity / Quasi-stationary

voiced — periodicity ⊕nt more
unvoiced — " less

Q How to check wether signal in frame is periodic or not?

Ans. Auto correlation — formulat^n ($\sum_{n=1}^{N} s(n)s(n-$

Correlation of two signals is high when the signals are same.
In Auto correlation,

if we apply auto correlation on signal (on 20 ms frame) then the correlation coefficient — {2nd peak & 1st peak} ∴ almost similar i.e correlat^n b/w the two in voiced
because of periodicity

but in unvoiced — there will be no peaks in auto correlat^n because no correlat^n b/w signal

deviat^n b/w first peak & second peak will be much higher

→ can use spectral energy in frequency domain and autocorrelation in time domain.

→ Excitation for unvoiced is like white noise so ideally, autocorrelatn should be an impulse func" for that, but these are few informatn that is left in unvoiced — hence we won't get an impulse.

→ Sometimes even noise can have periodicity so method ③ & ④ may fail these — so we need better methods for voiced/unvoiced detection.

How to handle them?

→ LP analysis

Q Can we use epoch information (→ glottal closure instants) is voiced/unvoiced detection?

Ans Locations & strengths of epochs will be periodic & high/good in case of voiced speech

& random for unvoiced speech. (rout pattern)

→ also a good approach even in case of noise (∵ no epochs in noise)

Teacher's Signature

Lecture 07 (1st September)

→ **Pitch**: Rate of vocal fold vibration
→ Impulse like excitations in voiced sounds
→ The periodicity of those impulses is pitch period. = $1/T \simeq$ 5ms to 9ms.

→ Children > Female > Male
(7-9) (5-7)

→ No concept of pitch in silence or unvoiced region.

Q Approaches to detect pitch from speech waveform

① Zero crossing — very erroneous
— won't work if noise is present

② Autocorrelation

③ Epoch locations using

→ successive differences of epochs will be pitch period

④ Cepstrum

Teacher's Signature

Q. How can we use autocorrelation for pitch detection?

Ans a. Segment size = 20ms - 50ms (Sampling rate ↓ 8 samples per ms) 8 Hz

b. ~~Difference~~ ~~Distance~~ b/w first two successive peaks (dominant) after autocorrelation is applied is the pitch period.

Q. Why does it give pitch?

Ans When the moving frame comes in same configuration i.e. when the cycles match — the correlation will be higher.

Cons: getting second dominant peak can be difficult in case of large no. of spurious peaks. if the segment size is not taken carefully

→ We get pitch only in voiced relation
→ You can put a threshold on the Ratio of first & second dominant peaks to determine voiced/unvoiced.

Ratio is less — voiced
" " more — unvoiced

Teacher's Signature

Short Term Fourier Transform Analysis

Need for STFT

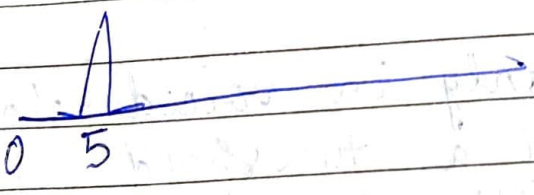
Q Signal = x Hz, FT will give

Ans an impulse at x. i.e. it indicates the frequency of the signal. So, the Fourier transform spectrum gives the frequency of the signal

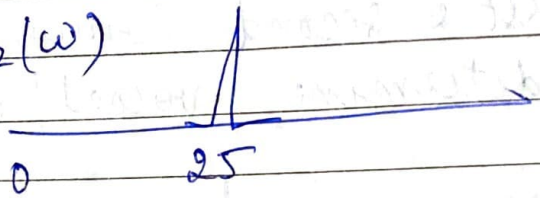
Stationary signal → do not change in time

eg
a. $x_1(t) = \cos(2\pi \cdot 5 \cdot t)$ — 5 Hz
b. $x_2(t) = \cos(2\pi \cdot 25 \cdot t)$ — 25 Hz
c. $x_3(t) = \cos(2\pi \cdot 50 \cdot t)$ — 50 Hz
d. $x_4(t) = x_1(t) + x_2(t) + x_3(t)$

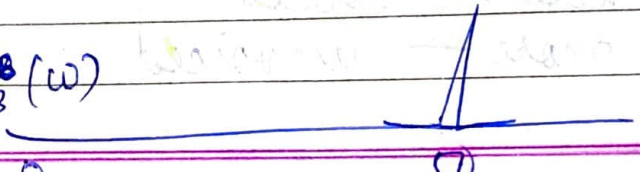
a'. $X_1(\omega) = FT(x_1(t))$



b'. $X_2(\omega)$



c'. $X_3(\omega)$

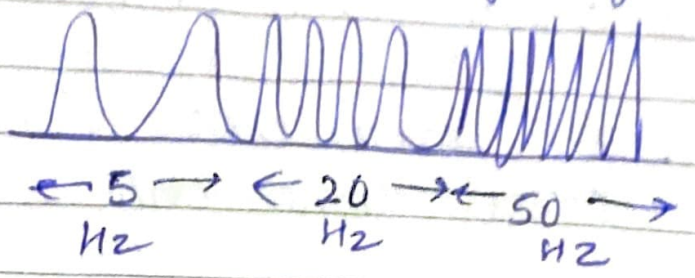


d. $X_4(\omega)$

non-st

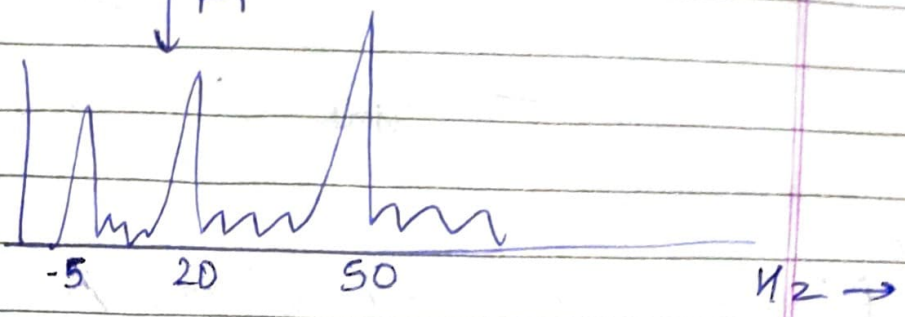
→ if we take a non-stationary signal

e. $x_5(t)$:



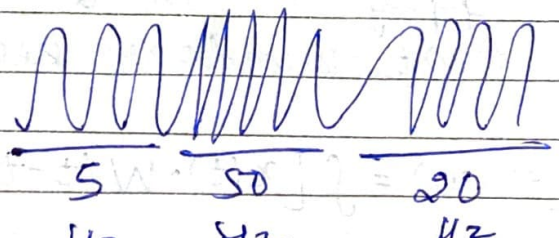
FT

$X_5(\omega)$:



So, FT gives us knowledge of what frequencies exist but no informat" about where these frequencies are located in time.

So, for eg. if we change $x_5(t)$ to $x_6(t)$ →



FT

will give same $X_5(\omega)$ spectrum

Drawback:

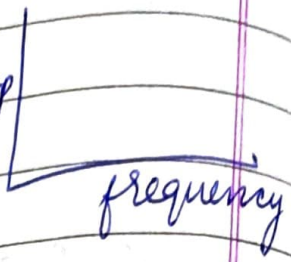
So, we cannot classify/distinguish b/w non-stationary signals using FT. Since for both $x_5(t)$ & $x_6(t)$ it gives same spectrum.

FT
↳ converts time to frequency

Pros: all frequency info is known
Cons: loses time info

Signal:

FT: amp



So,
FT works well if signal is stationary and speech is non-stationary so I need both time & frequency analysis

→ So, we use STFT — FT for short segments and find spectrogram i.e. time vs frequency plot/ spectrum.

→ In STFT, we apply FT on a windowed signal and that window keeps on moving

$$STFT_x^w(t', w) = \int_t [x(t) \cdot W(t-t')]\, e^{-jwt}\, dt$$

→ For $x_5(t)$ there will be only one magnitude spectrum if we apply FT.

But, in STFT (t vs Hz) — the spectrogram will vary because it depends on the window size

Teacher's Signature

Q Ideal length of window = 20ms (160 samples)

Q How to choose the ideal window length?

Ans:
Since FT works well for stationary signals
— within the window the signal should be stationary. So we take 20ms window size — since we assume the signal is stationary.

advantage STFT

Disadv. of STFT
└ a kind of uncertainty / tradeoff is present b/w time & frequency, since the two are opposite to each other. i.e. we have time resolutn (when time info is seen well) and frequency resolution (where frequencies well seen)

→ for non-stationary signals we require Joint time-frequency analysis

→ Output of STFT is spectrogram and it is not unique — changes ∝
① window size
② window shape ┬ Hanning (bell shape)
└ Hamming (bell shape)

Hamming (bell shape) can be changed in spectrum section in wavesurfer

PAGE NO.:
DATE: / /

→ The output of STFT that comes as a spectrogram is the convolution of signal spectrum and window spectrum because we are windowing the signal

→ Window Shape

→ bell shape window → these windows have less sidelobe leakage. Hence, the window won't affect much of the signal spectrum

→ but if we take rectangle window the spectrum of signal gets affected.

Q If we take a small window, which one is good - time resolution or frequency resolution and why?

Ans. Time Resolution → when we take narrow window

Frequency Resolutn - wide window because we will get all frequencies

The spectrogram flattens for high window size and does not capture any time variations

Teacher's Signature

$2^x = 32768$ $x =$

and for small window size — the spectrogram shows more time variations vs freq.

Q To get frequency info in FT what transform should be used?

~~N point DFT~~ n-DFT (Discrete FT) where 'n' should be in powers of 2.

→ To compute DFT fast we use FFT.

So n should be a value in powers of 2 greater than the no. of samples

→ n is the number of points where frequency values are computed and those points are joined

if n < no. of samples — we loose information

So for 20 ms frame & 8 kHz sampling rate FFT points should be 216 point or 512 point DFT.
(min) DFT

2^9 64 128 21

(DFT is applying FT on discrete signals and FFT is used to perform Fast DFT.)

→ Spectrogram is 3D plot —
Time vs Frequency vs Amplitute (Intensity)

→ We can also use spectrogram for Voiced/unvoiced detectn —

ratio of $\frac{\text{high freq}}{\text{low freq.}}$

more for unvoiced (∵ dominated by ↑ freq)
less for voiced (∵ dominated by ↓ freq)

→ formants tell us where spectrogram is dark

→ So in 512 point DFT all points other than the 160 samples are marked zero — this is called zero-padding

Q ≡ How does changing the window size affect the joint time frequency resolutn

if large window — same Spectrum.
you can see all freq. ⊕ nt but not when i.e. the time info.

so FR↑
TR↓

Teacher's Signature

Aliasing

when aliasing occurs — i.e. when there is an overlap in signal low frequencies will get affect which are the voiced regions.

(11) In discrete domain

output = convulatn (input, system response)
= $x[n] * h[n]$

Lecture 8 (~~18 Aug~~ 4th Sept) — STFT for ~~short~~ time scale modification

- if we resample —
  - (a) — discard 1/2 sample for every 5 sample
    - pitch cycle ↓es
    - → pitch ↑es
  - (b) undersampling — ~~increase~~ add zeros in b/w
    - pitch cycle increases
    - hence pitch ↓es

Q What should be window size?
if we do resampling in frequency domain — why pitch stays same

Resampling in time domain — we take sample acc. to time

PAGE NO.:
DATE: / /

However,

in spectrogram — resampling happens only in 20 ms.

So we have x 20ms frames

So we discard ~~con~~ / add frames

In 20 ms — 2-3 pitch cycles so the pitch is preserved and speaker charach. also.

Q why is it of Phase Vocoder

→ Phase Vocoder is used for time scale & pitch modificat^n
→ increasing/↓ sampling rate — will make the audio fast / slow.

→ Talking time — speeding up (sample it & in time domain)
↓
discard some samples

slow down — undersampling

Teacher's Signature

Quiz

Q1. Wider window TR ↑ — False (+1)

Q2. Speech Coding for cellular — True (+1)

Q3. Female pitch and bird — False (+1)

Q4. Synthesiser — 0 (+1)

Q5. Sampling Theorem → (-1/3)
L limits the min. freqy requirement.

Q6. 270 kbps (-1/3)

$8 - \frac{2}{3}$

$8 -$

7.03/10

Teacher's Signature

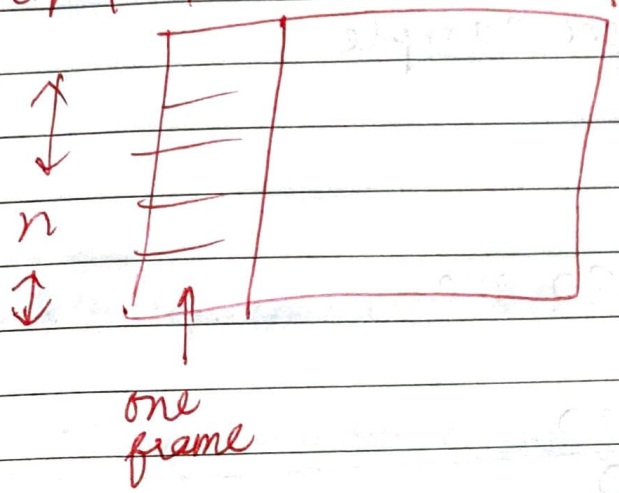
SSP A2

buffer — x into frames of size 'n'

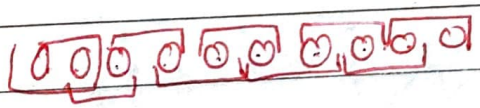
↳ x.size() - total #samples

n → # samples in a frame

output $\lceil L/n \rceil$ = no. of frames



94274 — 160 window size



(10-2+1)

94274 - 160 + 1

= 94115

94274 —

for i=0 to i< 94274 ; i+=2

1 sec 8000 samples

? 94274 sam

94274 / 8000 sec

Teacher's Signature

$8 \times 0.000125 \times x = 5.30$

20msec in one frame

$\frac{1}{8000}$ time per sample

$\frac{20 \times 8000}{1000} \times x = 5.30$ sec

$= \frac{530}{2}$

11.8

Teacher's Signature

Q Circular convolution for DFT. Why?

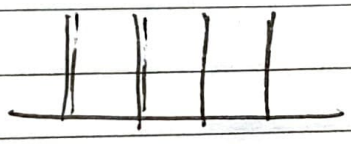
20 — 160
30 — 240

3×1

100

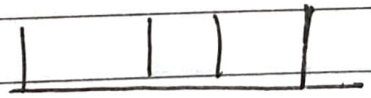
Class : 8th September 2020

Linear Time Invariant System

Impulse Train 

$x(t)$  (continuous signal)

↓ sample by multiplying with impulse train

$x(n)$  (discrete signal)

$x(n)$

Linearity condition

$$a x_1(t) + b(x_2(t)) \rightarrow a y_1(t) + b(y_2(t))$$

if $x_1 \rightarrow y_1$

$x_2 \rightarrow y_2$

Linear

Teacher's Signature

$y = mx$ ✓ linear

$y = mx^2$ ✗

(?) $y = mx + c$ ✗

Time Invariance

$x(t) \rightarrow y(t)$

$x(t-1) \rightarrow y(t-1)$

if input

Class 11th September

→ Z-transform

→ Impulse response & Discrete signals are represented in the form of z-transform.

→ Motivat^n: very easy for Discrete system analysis and also for some of the signals FT may not exist by z-transform exists. Why?

We can represent 'z' as $Ae^{j\phi}$

if $A = 1$, then it is DTFT only.

because even $x[n]$ not stable

by multiplying it into A^{-n}.

this - makes it as a stable signal

will become $A^{-j\omega N}$ ~~will become~~.

and then we can get $x[z]$

Q Given a signal, how to get z-transform

Teacher's Signature

Q Circular convolution for DFT. Why?

~~→ Signal is periodic~~

→ We discretize the input signal in time domain — ~~because of convolution property~~

HOW?

→ discretizing by taking impulse train and multiplying ∑ signal.

→ because of this multiplication — the spectrum is the convolution of the impulse response x signal response.

because of which the output becomes periodic and to avoid aliasing we want

$$f_s \geq 2 f_m$$

But,
Even after sampling in time domain, we cannot compute DFT on computer (because for that we need to discretize in frequency domain).

Now since the signal is periodic we cannot apply linear convolution so we apply circular convolution.

At the same time, we have to take 2 signals (constraint in circular conv. — should be of same length) — & apply convolution

PAGE NO.:
DATE: / /
SHREE

Teacher's Signature

→ In n-point DFT

if n > also - we use 0-padding technique (adding zeros to make same length/ n point)

✓ is

↓

this increases frequency resolution

Mel-Frequency / Cepstral Analysis

~~Spectrogram~~

→ Spectral features extracted using segmental analysis

→ what do these features represent?

→ modifying spectral features ___ into ___ cepstral features

~~Spectro~~

SPECTRUM ⟶ CEPSTRUM

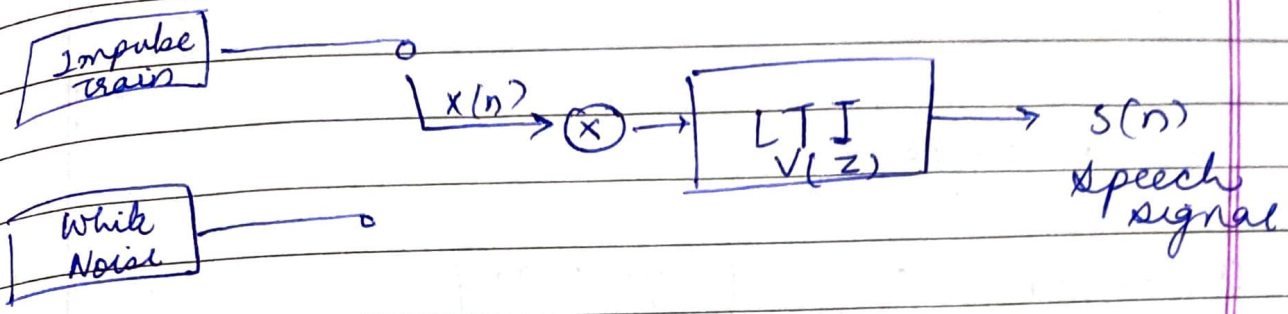
Cepstrum is inverse of Spectrum

PAGE NO.:
DATE: / /
SNEE

Teacher's Signature

Spectrogram

LTI



$$S(n) = x(n) \text{ convolut}^n V(n)$$

So, if we observe, we want information related to the vocal tract because it is responsible for changing sounds & it is unique for every human & it is the one that acts as resonator — hence it is required / more useful.

Unfortunately, the signal we receive is $S(n)$ which is convolution of $x(n)$ & $V(n)$

So we need to seperate the two — using LP analysis

$$\text{Cepstrum} = \text{Inverse FFT}(\log(\text{spectrum}))$$

11, 14, 17

SHREE
DATE : / /
PAGE NO. :

Teacher's Signature

$$FFT(s(k)) = x(k) * v(k)$$

$$\Rightarrow \log(x(k) * v(k))$$

$$\log[FFT(s(k))] = \log(x(k)) + \log(v(k))$$

Cepstral Analysis

after FFT — frequency

~~peaks~~

spectral env. peaks — formants

spectral env — spectrum of LP quotients

smooth spectrum of FFT.

How to achieve separation?

of $x(n) * v(n)$

i.e. given

$\log X[k]$ get $\log H[k]$, $\log E[k]$

$$\underbrace{\log X[k]}_{\text{spectrum}} = \underbrace{\log H[k]}_{\text{env. elope}} + \underbrace{\log E[k]}_{\text{details}}$$

PAGE NO.:
DATE: / /

Teacher's Signature

so IFFT (log X[k])

will separate the 2 into

high freq & low freq

1 1

details (excitation feat) envelope (VT feat)

log H[k] log E[k]

* first 13/15/17 values coeff out of 256 represent vocal tract

-- diff for each frame in these 13 values show how vocal tract is changing

Mel-frequency Analysis

mel-scale ← linear scale

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

Spectrum → mel-filters → Cepstral analysis

1

MFCC

Teacher's Signature

LP Analysis (15-09) — exploits redundancy of speech signal

$$\sum_{i=0}^{p} a(i)\, y(n-i) = \sum_{j=0}^{q} b(j)\, x(n-j)$$

Apply z-transform,

$$\frac{N(z)}{D(z)} = H(z) = \frac{\sum_{j=0}^{q} b(j) z^{-1}}{\sum_{i=0}^{p} a(i) z^{-1}}$$

numerator → roots of num. → zeros

denominator → roots of denom. → poles

When only numerator — all-zero system or Finite Impulse Response system or moving average (MA)

└ D(z) is unity

When $\frac{1}{\text{Denom}}$ = all-pole system Infinite Impulse response (Auto agressive) AR

└ N(z) — constant

vocal tract

When both — ARMA

Teacher's Signature

synthesis filter

speech $s(n)$ → [All zero filter] → residual $r(n)$ → [All pole] → speech

Analysis filter

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

In AR Model,

$$\hat{y}(n) = -\sum_{i=1}^{p} a(i)\, y(n-i)$$

(coeff for VTS vocal tract LT I)

(represents excitation)

$$e(n) = y(n) - \hat{y}(n)$$

$$e(n) = y_n - \sum_{i=1}^{p} a(i)\, y(n-i) \quad \text{---} (2)$$

Solve (2) by auto correlation approach

Total error =

$$E = \sum_{n=-\infty}^{\infty} e_n^2$$

$$\frac{\partial E}{\partial a(i)} = 0_i \qquad 1 \le i \le p$$

$$\Rightarrow \sum_{i=1}^{p} a_i\, r(k-i) = r(k) \qquad i \le k \le p$$

where $r(k) = \sum_{n=k}^{N-1} s_n s_{n-k}$

DATE: / / PAGE NO.:

So,

$$\begin{pmatrix} r_0 & r_1 & r_2 & \cdots & r(p-1) \\ r_1 & r_0 & r_1 & & r(p-2) \\ r_2 & r_1 & r_0 & \cdots & \\ \vdots & & & & \\ r(p-1) & & & & r(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} \\ \\ \\ \\ \end{pmatrix}$$

R — autocorrelation matrix

A

Solve

$$RA = R'$$

$$A = R^{-1} R'$$

A → LP coeff

Prediction Order

└ no. of prev values you should take depend on analysis bandwidth

$$p = \frac{2(BW)}{1000} + C$$

→ 1 foemant per kHz of BW

and

1 foemant requires 2 complex conjugate poles

· So, for every foemant 2 predictor coeff. → 2 ~~foemant~~ coeff per kHz of BW

Teacher's Signature

r_1
r_2
|
|
|
r_p
R'

greater P → more formants

V/UV
- Error more where more excitat^n
- errors - more for unvoiced

formant extraction
→ LP ~~residual~~ spectrum → formants
~~LP residual~~ → VT features (a_k)
error gives → excitat^n features

a_k - LPCC -

Pitch
Auto correlat^n of LP residual
↓
diff b/w 2 peaks.

PAGE NO.:
DATE: / /
SHREE

Q To get 5 formant for LP spectrum what should be LP order?

1 formant – 2 complex conjugates
⇒ 2 order

$5 \times 2 = 10$

Q Why cepstrum?
L to separate VT from excitat^n

Q Why LP analysis spectrum is smooth?

VT as all pole resonator
assuming are 0 resonator has only 5 responses (P/2) – so accordingly only 5 frequency are highlighted.

We get spectrum from LP coeff not directly from signal
↓
we removed excitat^n
so no zig-zag only
vocal tract is highlighted

Teacher's Signature

Q Is MFCC enough for ML model?

yes.

Q In which scenario MFCC may not be enough.

eg. emotional recognition because it requires excitat" features like strength of excitat"

Q If training & testing condition is not matched
↳ then LPCC is better or MFCC — ans: MFCC

Q LP Coeff
↳ using autocorrelat"
↓
minimizing the diff of eq" of original & predicted signal

Q Inverse filter is all pole filter.

No / false

Q Residual → speech } all pole filter.

Teacher's Signature

25th September - Epoch Extract"

Method

① HE (LP Residual) - Peaks

② ZFF - directly from speach signal

doesn't uses these concepts

- ① Instantaneous frequency which comes from analytical signal
- ② Property of Impulse (~~Epoch local~~)

Q What is FFT (impulse)

- constant (DC signal)

impulse will have influence on all frequencies

→ Instantataneous frequency represents a change whereever there is an impulse

→ filter signal at diff frequency to get plot.

(depends on VT influence on diff frequences

so if VT influence less on 500Hz

Plot will show epoch local" better

25, 29, 30

Teacher's Signature

Because impulse will have influence on all freq but VT can have on some.

Since VT doesn't influence low frequency (it start at 100 Hz) etc) so choose / filter signal at low freq.

ZFF

→ ZFF is a low pass filter (at low Hz)
→ Ideal ZFF is incubator i.e. add the samples i.e. integrating

(a) Difference signal to remove time varying low freq. bias i.e. pre-emphasis

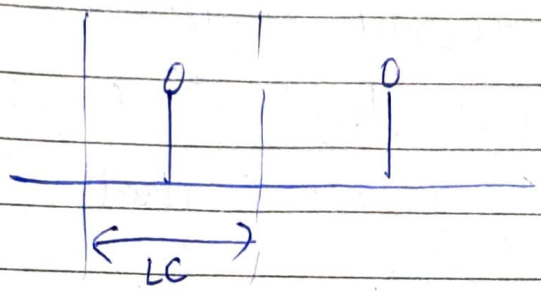
→ 2 times ZFF = 4 times integrator

(2) Trend Removal
Remove mean from each value around M values.

(3) negative to positive zero-crossing gives the epoch locatn in mean sub signal (mss)

Evaluat^n approach

Larynx Cycle (LC) — 1/2 pitch period on both sides of epoch



So within larynx cycle if nothing is ⊕nt ⇒ epoch locat^n missed
or if is more than one
⇒ false alarm.

when only one in LC — ✓

→ Can be used for V and UV detect^n.

---

all peaks of impulse not same as strength of excitat^n is diff.

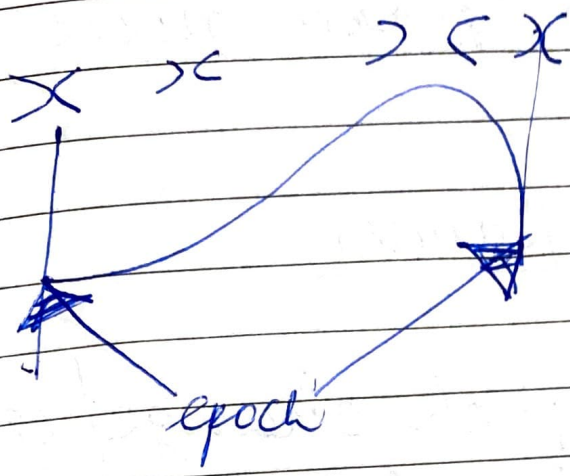
↓

is given by slope of MSS
sharply — strength ↑
slow slope — low strength

Teacher's Signature

GVV - Glottal Volume Velocity

Excitation feature



How to get GVV

LP residual → [Integrator] → GVV

→ How fast opening/closing vocal folds

→ useful for emotion, singing speech

→ works well even if noise

→ CMU Arctic Database

Teacher's Signature

Noise

- reverberat"
- bg noise
- babble / multi-speaker noise

(29th-09) Prosody features

↳ suprasegmental

① Intonat" pattn → pitch change / contour
② Duration → seq length of syllables – rate of speaking
③ Intensity → energy contour

② Durat" is difficult to extract

Issues in prosody extract"

① what is high for someone may be normal for other

② for each ind. we need to know normal voice first to understand if there is a change

③ Durat" is difficult

PAGE NO.:
DATE: / /

Teacher's Signature

Applicatn

1. TTS
2. Lang. Identi
3. Emot^n
4. Speaker recog
5. Speech recog

→ extract for whole region and use them as anchor points

Teacher's Signature

6-9 Oct missing

~~1~~, ~~2~~, ~~3~~, ~~4~~, 27, 30

3, 6, 10

Teacher's Signature

13th October

VOP – Vowel onset point

- vowel regions – ROBUST feature ~~extraction~~

Indian L – syllabic
- L has vowel nuclei

helps in
- speech recognition
- prosody modificat^n
  - changing durat^n of speech
    - L ~~stretch~~ / squeeze only vowels

→ vowels act as anchor points

→ Vowel onset ← Vowel region → Vowel offset

→ 3 approaches :

complimentary to each other
- ~~LPCC~~ source = exitat^n source – LP residual
- spectral peaks (10 peak values) are summed
- time informat^n (modulat^n spectrum)
  - ↓
  - time domain envelop of signal

all three show significant change at VOP

PAGE NO.:
DATE: / /

Teacher's Signature

① Signal
② LP Residual
③ Hilbert Envelope of LP residual

assumptn – significant change in envelop – VOP – peaks

④ but spurious peaks
Eliminate – using – first order diff & keep a threshold

peak & valley values are normalised

⑤ convolutn ε first order Gaussian differentiator
L = 500 (length)
σ = 200 (sigma)

⑥ Peaks in the final signal are VOP

-ve peak – V offset point

Teacher's Signature

16th October (VOP)

- Spectral info is better for coded speech
- Changes =
  - instead of n peak values find epoch locations using ZSF
  - Spectral energy around ei is high
  - within 2 epochs, 30% of cycle & estimated spectrum — vowel/spectral energy will be high

AI & ML Trends

- Speaker, language, emotion — simple
- ASR, TTS — advanced
- data — diff accent
  diff gender, age group

PAGE NO.:

Teacher's Signature

GMM (Gaussian mixture Modelling)

- lot of data
- classes - 23 (languages)

(1) assume some gaussians and fit the features into these gaussians more gaussians is good because they capture more informat" provided we have enough data

(2) → GMM : universal background model where
VBM will be built : huge data
& adopt that model to the target data (1024 mixtures)

We are only interested in speaker dependent component of the VBM vector

(3) So use Joint factor analysis

convert M into $m + (T)(w)$ — intermediate vector

VBM Super vector → compressed into 200-300 → I vector

m → VBM

T → Total variability matrix

w → latent variables

I = T & w

Teacher's Signature

④ I-vectors are used as a model & also as a feature vector.
Then use SVM & DNN for classificatn

Now, I-vector has been replaced by DNN & attentn

DNN
- discriminative models at large
  - weights of connectn are optimised that is the model
- GMM is generative — for each class a diff GMM but here in DNN — only one network that classifies all classes.
- fail when training & testing conditions are diff
- work well with large data

- Architecture
- Layers
  - Neurons
  - learning Rate etc

} Parameters

PAGE NO.:
DATE: / /

Teacher's Signature

DNN & attentn

- Frame level decision
- Utterance level decision

eg Lang. IDent.
- 10 lgs
-

attn mechanism
L take decision at utterance level & not at frame level
↓
context vector

→ frames are given to attentn model & a weighted average of the hidden layers give context vector

→ So attentn weights are low in silence region

→ Multi-headed attentn

LID — explicit — STT - then find L
implicit — feature — classf

ASR (Automatic Speech Recognit)

- Speech to Text
- Language Model + Acoustic Model

Y - MFCC
W - sequence of words

STT (ASR) = $P(W|Y)$

$$P(W|Y) = \frac{P(Y|W)P(W)}{P(Y)}$$

$P(Y|W)$ → acM
$P(W)$ → lang M
$P(Y)$ → constant

AcM
- → wav files
- → transcriptn
- → pronunciatn dictionary — each word what phones are Ont
  - ★ comprehensive — all possible words
- → find phone boundaries

L multiple pronunc — mentioned (depending on context)

Teacher's Signature

20th October

A2

Problems =

① Machine listens phoneme — word — sentence

to errors may creep in

② Data — should nullify the accent / age group / gender / speaking style variation

③ Fragmentatn — chop wav files into fragments & manually transcri

LM — probability of sequence of phones reduces / cuts down the errors to predict text

A4

Evaluatn

Word Error Rate

① Substitutn
② Insertn — extra word
③ Deletn

$$\frac{S+I+D}{N} = \frac{S+D+I}{S+D+C}$$

SHREE PAGE NO.: DATE: / /

Earlier
Dynamic Time Working

Better — HMM (Hidden Marko Model)
A6 { data < 1000 hours
else HMM gets saturated

Acoustic Models

A5
① GMM - HMM (captures sequential info) (most popular)
② DNN - HMM

→ ASR are speaker & domain specific
→ 3 diff ASR can be modeled :
A7 - Read speech - reading
- Extempore - lecture/talks
- Conversational - phone / meeting
So one ASR may not work for all

HMM — model the temporal variability

A8 → each phoneme will have a HMM
$\lambda = (A, B, \pi)$

→ three states — beg
middle
end

→ depending on context, position & duration pronunciation of a phoneme changes

Teacher's Signature

Hidden states :

- we don't know where the phoneme is starting / ending
- assume 3 hidden states in the product^n of a phoneme

Markov Process

- what is the prob. that I am in a particular state for time $t=0$ & what is the prob that I will remain in the same state or go the ~~xxx~~ some other state.
  (state probabilities)

in speech HMM

- we cannot go back
- only left to right topology

A = state transit^n prob. ($a_{ij} \rightarrow i \rightarrow j$)

B = symbol probability ($b_{ik} = P(k^{th} \text{ sym at } i \text{ state})$)

↳ GMM

π = initial probability

Three Basic Problems

(1) Evaluation Problem

$P(O|\lambda)$ — given model & output

P (modelled produced output)

forward & Backward algo

(2) Decoding Problem — give m & o most likely state sequence

written-Bi

(3) Learning Problem — parameters to get high prob of generating sequences ∈ model

Limitation:

- limited parameters to optimise
- so will not change ∝ increasing data — cant optimise much

on the other hand - NN - have a lot of parameters

(19)

Teacher's Signature

Q Sequential Info is used in which applications :

a. LID (X)

b. Speech recog (✓)

c. Speaker / Text Independent sys (X)

Q What is A in HMM

A – matrix

What are elements of a_{ij}

Eg. Transition from i to j

prob

Q What is symbol prob.

B_{jk} – prob of symbol k in state j

N
↳ no. of states

M
↳ no. of symbols

optimised based on experiments

Teacher's Signature

Q What is input to AM
- MFCC

Q What is input to LM
- Transcripts

Q Transcript^n dict
- phonetic descript^n of each word

Q HMM vs DNN

HMM
- one HMM for one phoneme (generative)

DNN
- one model for entire classificat^n (discriminative)
- one model classifies all phonemes

A10

October 2[8]7

Teacher's Signature

Nov 3 (Speech Enhancement)

- once speech is degraded with noise - how to enhance speech & improve the quality so that intelligibility res.

Speaker
Speech Recognition

- Naturalness & intelligibility

Apl
- Cell Phone speech enhan
- Spch Spch R

3 kinds of degradation

1. Bg noise - AC noise $s(n) + d(n)$
2. Reverberat^n - Reverberant/reflected speech $s(n) * h(n)$
3. Competing Speaker - $s_1(n) + s_2(n)$ (interfering speech)
   - most difficult since both req.
   - x not req have the same characteristics

PAGE NO.:
DATE: / /

Teacher's Signature

Approach

Pros = effective

Problem — need for explicit modeling of degraded component

Prior ① Spectral Processing — estimate the noise spectrum & remove it from ~~degraded~~ speech spectrum

② Temporal processing — in time domain
└ enhances the high SNR region

└ Pros = may not require explicit modeling

└ Demerit: Ineffectiveness

LP residual ⇏ LPC → speech

Clean ≠ Clean → Clean

D ≠ D → D

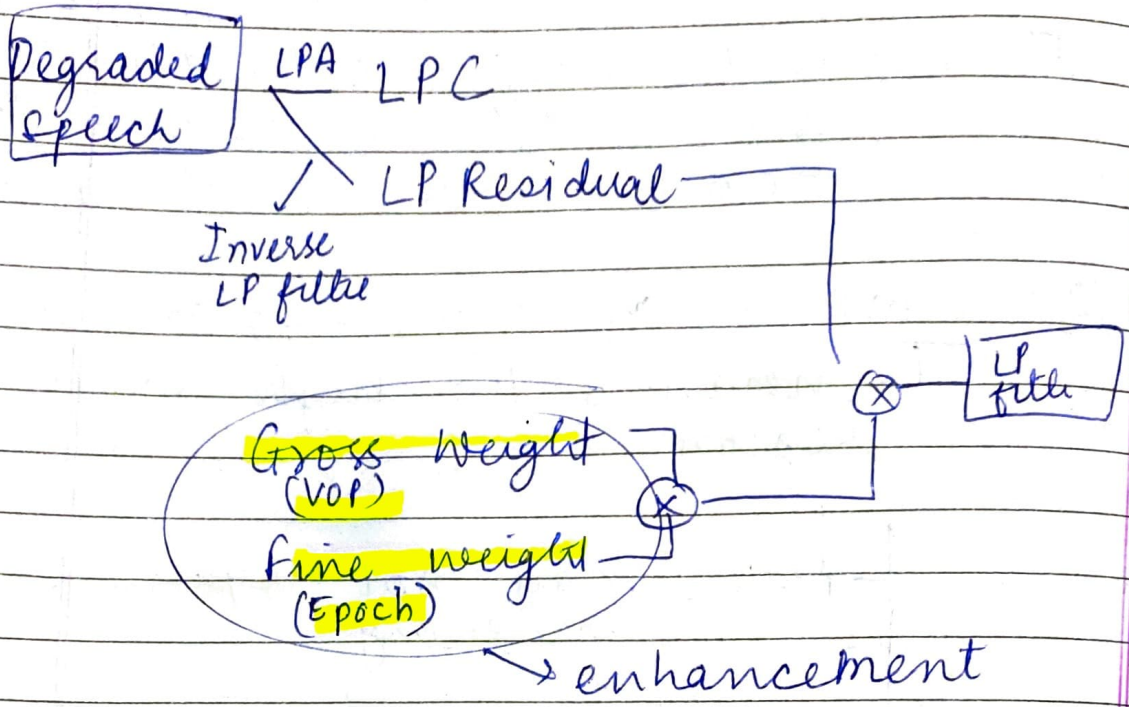
C ≠ D

So, the enhanced (LP residual) ← enhance
+ Degraded LPC

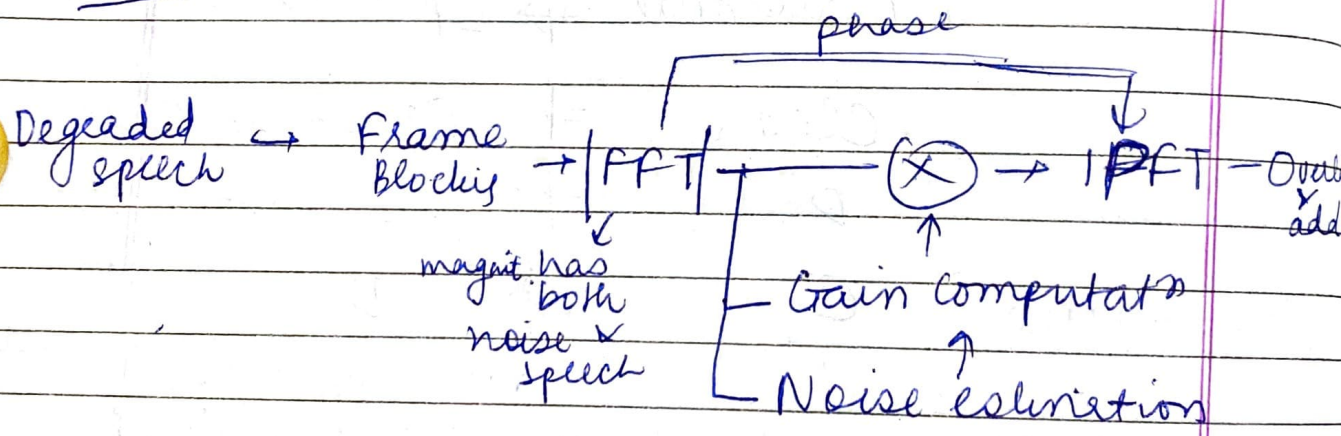
PAGE NO.:

DATE: / /

Temporal



Spectral



approach: Spectral subtract"

→ Diff noise have diff spectrum

Teacher's Signature

Music Noise

- Speech may be ⊕ in noise so when we do spectral subtraction — it creates some zeros in the spectrum

Temporal (non-speech regions become zero)

Gross Level (0-1 functn)

- get window funct^n
- plot 10 peak values (as in VOP detect)

So 3 evidences —

- ~~LP residual~~ sum of peaks in DFT spectrum
- Hilbert envelop
- modulatn spectrum

~~are combined to enhance~~

→ Non-linear Mapping funct^n on the ~~RMS~~ normalised sum

- making things zero in the low SNR (low energy region)

Teacher's Signature

Finite weight funct^n

- Find epoch using zero crossing 255 / peak of HE
- then take 10 values around epoch
- create a finite weight funct^n

Final

→ Combine Gross level f^n & finite weight f^n

→ Multiply LP residual & the two

Method: combinat^n of the two

- use enhance LP residual
- Temporally Processed speech is used to estimate the noise spectrum
- Noise estimat^n is done using (in spec proc) the noisy speech only and that will be applied on temporally enhanced speech

Teacher's Signature

Nov 6

TTS
- natural & intelligible
- need only one speaker data

Front End —— Back End

① Text Analysis
- Text Normalisat°
- Linguistic Analysis

Phonetic Analysis
Grapheme to Phoneme

- Prosodic Analysis
  - Pitch & Durat° (depending on content)

Back-End
- Speech synthesis
  - approach:
    concatenative synthesis
    (Unit select° synthesis)
    - phoneme level
      or
      syllable level
      ↓
      select right unit & concat-enate

Teacher's Signature

Optimal Text Select"

- collect valid sentences
- select sentences that cover all sounds of a lang. in no. of combinations & no. of times (in beg, middle, end) — all variants of sound

Problem — emotionless
- always need a backend database
- time consuming

Adv
- more natural & intelligible

Parametric Synthesis
- using HMM
- Problem — naturalness (less)
- save parameters instead of the database

Pros → reduce size
- lower development cost
- flexibility — modify parameters to generate
- so attractive — diff voices because
  - eg. storytelling

PAGE NO.:
DATE: / /

Teacher's Signature

Cons
- less intelligibility
- robotic - less natural
- speaker is fixed - same speaker to we get bored and less attractive
- concatenate % con

Wave Net

1. How to create emotional speech

TTS is Evaluated - Mean Opinion Score

Nov 10

Course

- Speech Basics
  - Product^n
  - LTI model of speech
  - Modelling
  - V/UV detect^n, pitch detect^n
- features Ex
  - Spectral - seg
  - Excitat^n - subseg
  - Prosody - supra
- Speech Applicat^n

NPTEL

STS (Bahubhashak) - reduce lg barrier

- directly - ~~ASR~~
- ASR + TTS
  - ↓
  - ~~loss~~ - emotion/ ~~charcteristics~~/prosody
  - NPTEL is naturally recorded
    - Eg. - filling words (pet), correct^n, repetition

Teacher's Signature

Problems ε Indian L

1. Low Resource
2. Morphologically rich
3. ~~Seq~~ free-order language

NN may help when structure is same from source L to target L

So we follow ASR + TTS.

1. ASR (Word Error Rate)
2. MT (fragmentat^n effects MT) so post processing is req.
   - BLEU Score (not every efficient - if the sentence is carrying meaning)
   - correct of output
   - sentence boundary
3. TTS
   - naturalness
   - intelligibility

Problems ε ASR

- Time alignment
- lip syncing
- use same filler words?
- use incorrect text also
- one L - less words
  other L - more words