SpeechTech

Week-1

L1 Course Introduct^n

- basics → speech, DSP, stats
- ASR → HMM, DNN, TDNN, LM
- ASR DL → E2E, RNN, Transformers, LLM
- Self supervised models → BERT, GPT, APC, Wave2Vec, HuBert
- Speaker verifica^n → x, e, s vectors
- TTS tasks
- primary purpose → communica^n. Formulate the msg. in brain, then lang. & sentence form^n is selected. The mouth articulators appropriate movement & finally spoken audio comes.

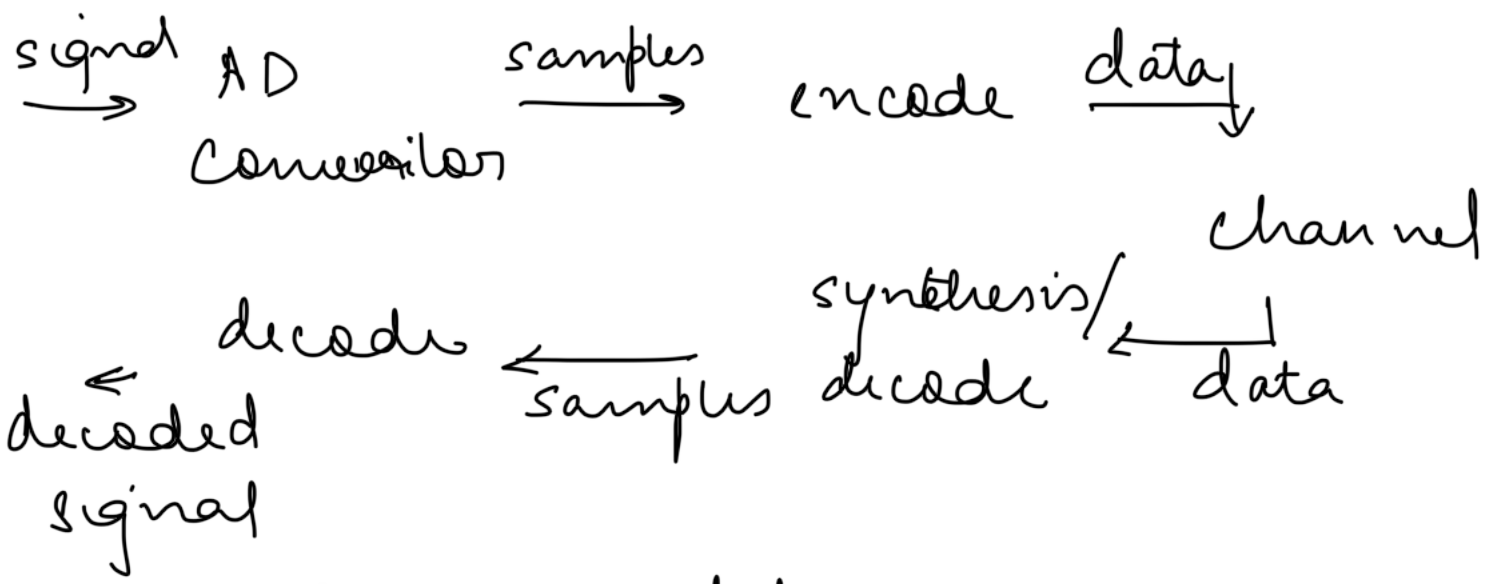
- Message → Lang. Code → Neuro-Muscular control → Vocal Tract → Transmission Channel → Basilar membran

Speech produc^n

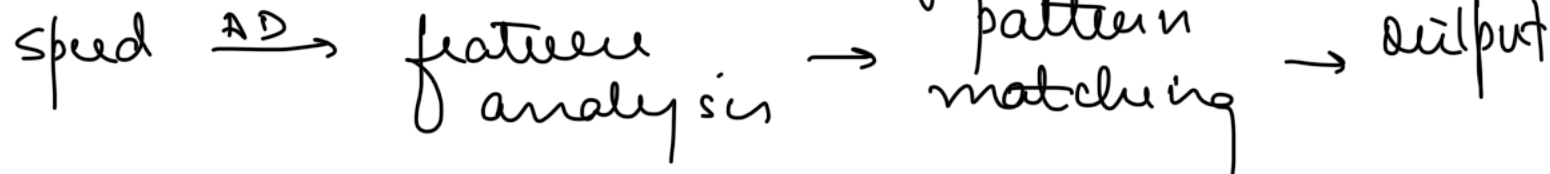
Speech percep^n

Basilar membran → Neural Transduc^n → Lang. Transla^n → Message Under-standing

- Speaker → Linguistic, Physiological, Acoustic
- Listener → Physiological, Linguistic
- App. I → Store & transmit speech



- TTS, Pattern matching



- DSP Applicaⁿ — STT, TTS, Diarizaⁿ, Speech synthesis, Verificaⁿ, quality enhancement, Aids for differently abled.

L2 DSP Fundamentals

- Sampling → nos. only can be handled. Therefore, A-D (analog to digital conversion). They are equally spa discrete pts.

- Sampling Th. → Under conditions, we can reconstruct the original cont. time signals from samples.
- By Nyquist - Shannon sampling th.
  Sampling Freq. ≥ 2 × Max. freq. in signal!
- Incorrect sampling freq. → loss of info. (aliasing)
  sampled signal look like a resultant of a slowly varying signal.
- The discrete sampled signals are then quantised (to a finite precision) to get the digital signals.

-

Signals

Narrow Band
→ telephone
→ channel acts as band pass filter (300 to 3.4 kHz)
→ sampling freq. = 8 kHz
  ↳ 8 k samples every second.

Wide Band
→ current std.
→ HD Voice (50 - 7 kHz)
→ 16 kHz
  ↓
  16 k samples every second.

- Speech
  - Telephone → A-law, mu law (64 kbits/s)
  - Parametric → based on code-excited linear predic^n (mobile)

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

  - GSM & CDMA → 7–13 kbits/s
- Audio – 44.1 kHz & 48 kHz
- MP3 – are lossy & operate at lower bit rates but are still CD quality
- we are sensitive to frequencies → eg. color (rods & cones) & speech.
- Recall –

$$\hat{i} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \hat{j} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \hat{k} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 4\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 2\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

these are the basis func^n.
Any pt. in 3D can be expressed as linear combina^n of unit vectors.

- A signal can be expressed as a linear combinan of complex sinusoids of diff. freq. ⇒ Fourier series

$$f(x) = \sum_{-\infty}^{\infty} c_n e^{\frac{inx}{L}}$$

↳ Transformaⁿ from time-variable space to freq. & variable space. 2 diff. views of same signal.

- Any periodic signal can be represented as a sum of complex sinusoids with diff. freq. Both conti. & discrete is known.

- CTFS

$$f(x) = \sum_{-\infty}^{\infty} c_n e^{\frac{inx}{L}}$$

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx}$$

→ Input - Conti. & Periodic
→ Output - Discrete & Aperiodic

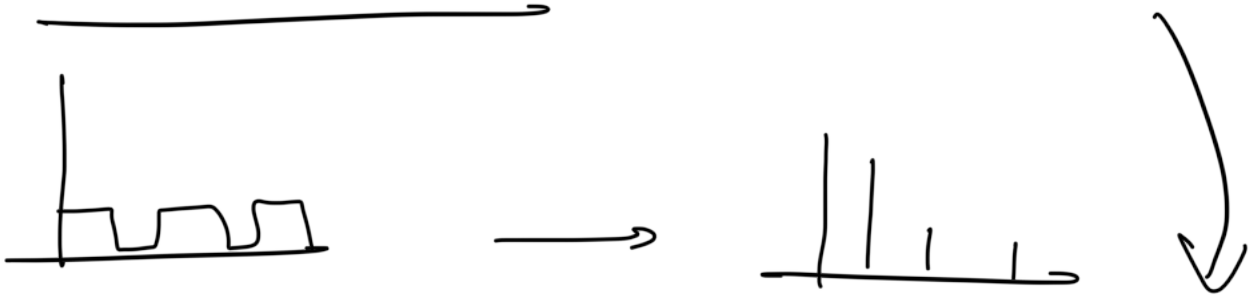
- DTFS

$$f([n]) = \sum_{k=0}^{N-1} c_k e^{i w_0 k n}$$

$$a^k = \frac{1}{N} \sum_{n=0}^{N-1} x(n)\, e^{-\frac{i 2\pi k n}{N}}$$

Discrete & Periodic both sides

L3 Fourier Series



- DTFS leads to DFT which is usually the transform we use when analysing on computers.
- CTFT

$$X(w) = \int_{-\infty}^{\infty} x(t)\, e^{-iwt}\, dt$$

I O → Continuous & Aperiodic

- DTFT

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

I → Discrete & Aperiodic

O → Continuous & Periodic

- DFT

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{\frac{-i 2\pi k n}{N}}$$

I/O → discrete & Periodic

used in analysis of freq. components in a signal. Used in actual DSP with computers.

- Z-transform → system manipulates the signals. This is a tool for understanding the interactn btw. signals & systems. For discrete. A signal w/o DFT FFT can have Z-transform.

$$X(z) = \sum_{-\infty}^{\infty} x(n)\, z^{-n}$$

- Sys. $x[n] \xrightarrow[\text{sys.}]{} y[n]$

Properties

→ linear / non

→ time variant / non

→ memory / less

→ causal / anti - / non -

→ stable / unstable

- Linear

$x_1[n] \xrightarrow{sys} y_1[n]$ $\quad$ $x_2[n] \xrightarrow{sys.} y_2[n]$

$x_1[n] + x_2[n] \xrightarrow{sys.} y_1[n] + y_2[n]$

- Time Variant

$x[n] \xrightarrow{sys.} y[n]$

$x[n-k] \xrightarrow{sys.} y_2[n-k]$

- memoryless → $y[n]$ depends only on the current sample of $x[n]$.
- causal → present & past $x[n]$

  anti → present & future $x[n]$

  non. → present, past, future $x[n]$
- Stable → $y[n]$ is bounded for $x[n]$

  BIBO
- LTI (linear & time variant) sys.

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

↓ convolution operation

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

Time domain → convolⁿ

Freq. domain → multiplicaⁿ

filters to remove the noise & get the voiced signal in a better way.

- Rational Digital filters

FIR — Finite Impulse Response

$a_k = 0$

→ does not take past outputs into accnt.

→ causal (takes current & past inputs)

IIR — Infinite Impulse Response

$b_n = 0$

→ takes past inputs & outputs into account

- Filter diff. eqⁿ

$$y[n] - \sum_{k=1}^{N} a_k y[n-k] = \sum_{n=1}^{M} b_n x[n-q]$$

- Filter transfer funcⁿ

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{n=0}^{M} b_n z^{-n}}{1 - \sum_{k=1}^{N} a_k z^{-k}}$$

$$H(z) = k \frac{(z - z_1)(z - z_2)}{(z - p_1)(z - p_2)(z - p_3)}$$

$z_1, z_2 \rightarrow$ zeroes

poles of den. $\rightarrow$ poles of filter

stable $\rightarrow$ ROC must be within unit circle.

Causal $\rightarrow$ ROC must be outside of the outermost pole.

stable & causal $\rightarrow$ all poles must be within unit circle.

L4 Speech Production

- Lungs $\rightarrow$ Trachea $\rightarrow$ Glottis $\rightarrow$ Nasal cavity + mouth + pharynx

excitation source — air flow path — open / close $\downarrow$ quasi-periodically

- time period of vibration cycle depends on the height & weight.

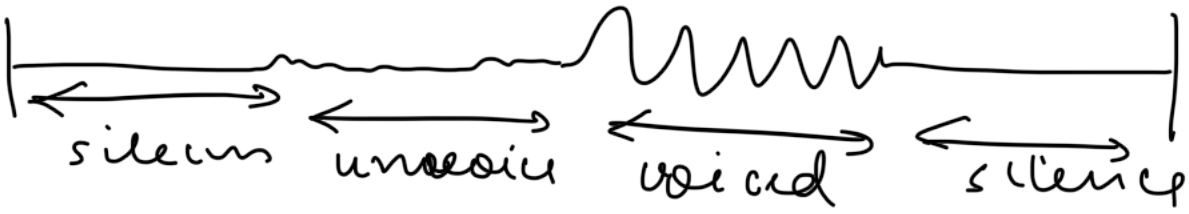
$T_{men} > T_{women} > T_{children}$

- different mouth configurations leads to diff voices (vowels)

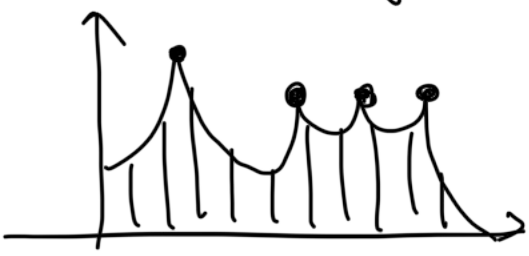
- all spoken vowels are voiced. phonemes are like alphabets for speech. They are the basic units from which speech is formed.
- Sound articulation → glide, nasal, stop, fricative, voicing, mixed source, whispered.

  labial, labiodental, dental, alveolar, palatal, velar, pharyngeal

L6 Waveforms



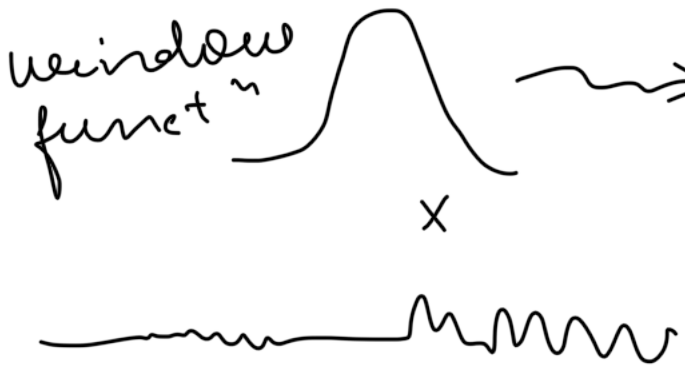
- Speech production model → mathematical equivalent of speech, also c/d as terminal analogues model



- formants ↓ marks the sound

| vertical lines ↓ pitch harmonics

- short time fourier transform (STFT)



→ window moves with some overlap

↓

Fake fourier transform

spectrogram → log magnitude of the resultant

wideband

→ small time window

→ good time resolu^n

→ captures formants

narrowband

→ larger time window

→ poor

→ good freq. resolu^n - pitch harmonics