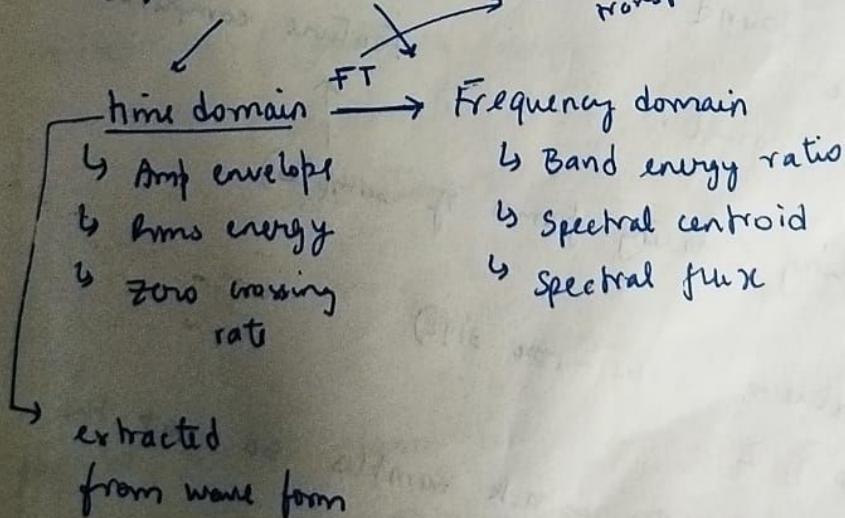
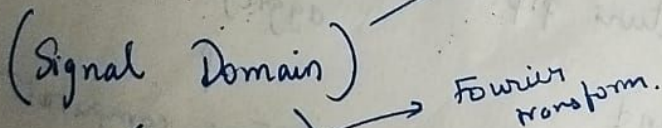
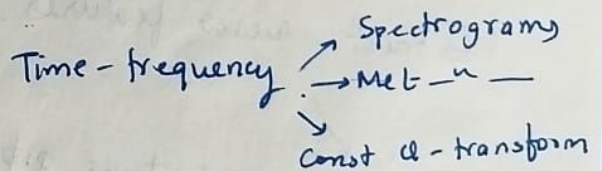
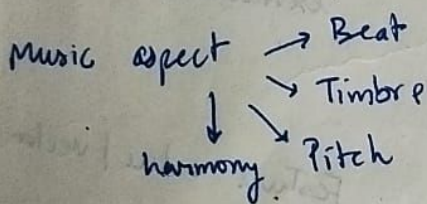
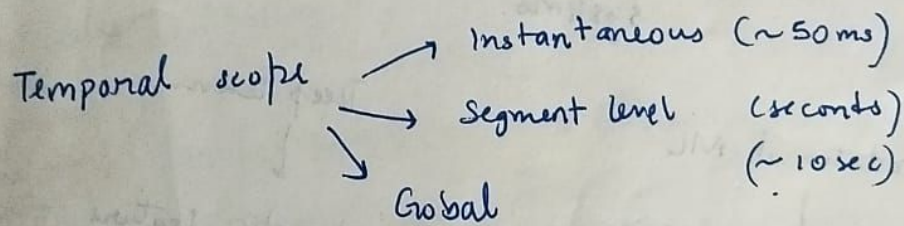
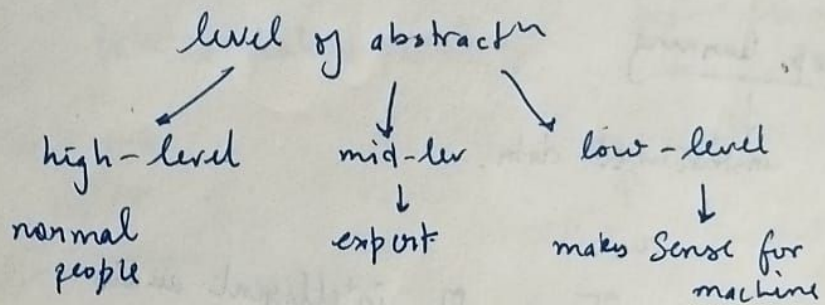


Audio features



Computing MFCC

Wave form



DFT



log-Amp spectrum



mel-scaling



Discrete cosine transform



MFCCs

Why Discrete cosine transform -

- Simplified version of FT
- Get real valued coeff.
- Decorrelate energy in different mel band
- Reduces # dimensions to represent spectrum



Gives the coeff.



Tells us how well that frequency fits in the signal

coeff ↑ ⇒ good fitting.

~~How~~ How many coeff?

• Traditionally 1st 12-13 coeff.

• Use Δ & $\Delta\Delta$ MFCCs

• Total 39 coeff per frame.

(derivatives) change in coeff in subsequent frames

Mel Spectrograms.

Humans perceive frequency logarithmically.

Mel-scale

$$1000 \text{ mel} = 1000 \text{ Hz}$$

$$m = 2595 \cdot \log\left(1 + \frac{f}{500}\right)$$

Mel-Spectrograms

- 1) STFT
- 2) Amp \rightarrow DBs
- 3) Convert frequency to Mel scale
 \downarrow
 1. Choose number of mel bands
 2. Construct mel filter banks
 3. Apply mel filter banks to spectrograms

Cepstrum



Spectrum

Quafrency

Frequency

Liftering

filtering

Rhamonic

harmonic

log spectrum.

spectrum

$C(x(t))$

↓
time-domain
signal

$$= F^{-1} \left[\log \left(\underbrace{F(x(t))}_{\text{DFT}} \right) \right]$$

Cepstrum

Speech = Convolution of vocal tract frequency response with glottal pulse.

Formalising speech

$$x(t) = e(t) \cdot h(t)$$

$$X(t) = E(t) \cdot H(t)$$

$$\log(X(t)) = \log(E(t)) + \log(H(t))$$

Speech

Glottal
Glottal

vocal tract
frequency Response

↓ IDFT

$$X(t) = \cancel{E(t)} + H(t)$$

not interested (low pass filter)

M-L approach $\begin{cases} \text{traditional} \\ \text{Deep learning.} \end{cases}$

Deep learning

↓
unstructured data.

Types of intelligent audio systems.

Traditional ML

↓
feature engineering

Deep learn

↓
automatic feature extraction

Extract audio features.

Time-domain feature pipeline

↓
Analogue sound

↓ ADC
digital signal.

↓
Framing (making frames of samples)

Feature value / vector

↑
aggregation

↑
Feature computation

Typical value \rightarrow (256 - ~~128~~ 8192)
powers of 2

Frames are constructed to stack samples so that the duration is humanly perceivable.

Fourier transform.

- Compare signal with sinusoids of various frequencies
- for each frequency we get a magnitude & a phase
- high mag indicates high similarity btw signal & a sinusoid.

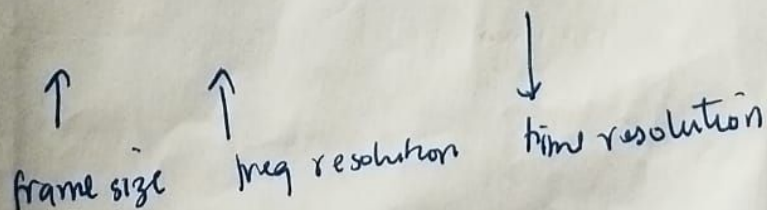
Short Fourier Transform.

- Apply FT for frames.

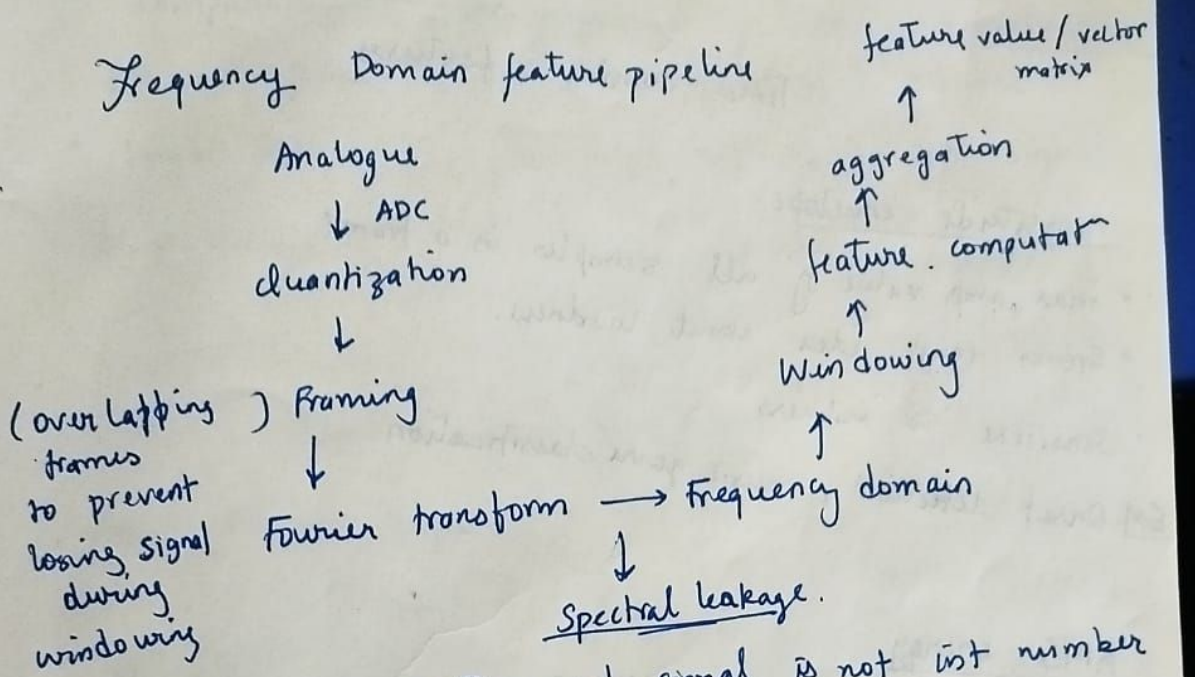
Outputs.

- DFT
 - spectral vector (# frequency bins)
 - N complex Fourier coeff
- STFT
 - spectral matrix (# freq bins, # frames)
 - complex Fourier coeff.

Time / frequency trade off.



Duration of 1 set sample \ll human resolution



- Processed signal is not int number of periods.
- End points are discontinuous.
- Discontinuities appear as high-freq components not present in the org signal

Windowing: { Applying windows function to each frame
eliminates samples at both ends of a frame
Generates a periodic signal.
↓
minimizes spectral leakage.

To make it periodic.

popular func: Hann window.

hop length: overlap betw ~~frames~~ frames.

Time domain features

Amplitude envelope.

- max amp value of all samples in a frame.
- Gives rough idea about loudness.
- Sensitive to outliers

Ex) Onset detection, music genre classification

RMS - energy.

- RMS of all samples in a frame.
- Indicator of loudness
- less sensitive to outliers than AE

Ex) Audio segmentation, music genre classification

Zero crossing rate

- number of times a signal crosses $x=0$
- Ex Recognizing ~~percussive~~ percussive vs pitched sounds

Speech

Recognition

Fourier

- Compa
- for each
- high
- ↳ a s

• Apply t

Output

- DFT
 - Sp
 - N

• STFT

- Sp
- Cor

Time / f



frame