Audio features

level of abstractⁿ

- high-level → normal people
- mid-lev → expert
- low-level → makes sense for machine

Temporal scope

- Instantaneous (~50ms)
- Segment level (seconds) (~10sec)
- Gobal

Music aspect

- Beat
- Timbre
- Pitch
- harmony

(Signal Domain)

- Time-frequency
  - Spectrogram
  - Mel-"-
  - Const Q-transform
- time domain $\xrightarrow{FT}$ Frequency domain (Fourier transform.)

time domain

- Amp envelope
- Rms energy
- Zero crossing rate

→ extracted from wave form

Frequency domain

- Band energy ratio
- Spectral centroid
- Spectral flux

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.

⇓

Gives the coeff.
↓
Tells us how well that frequency fits in the signal

coeff ↑ ⇒ good fitting.

- Decorrelate energy in different mel band
- Reduces # dimensions to represent spectrum

~~Home~~ How many coeff?

- Traditionally 1st 12-13 coeff.
- Use Δ & ΔΔ MFCCs (derivatives) change in coeff in subsequent frames
- Total 39 coeff per frame.

Mel Spectrograms.

Humans perceive frequency logarithmically.

Mel-scale

1000 mel = 1000 Hz

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

Mel-Spectrograms

1) STFT
2) Amp $\longrightarrow$ 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

Quefrency → Frequency

Liftering → filtering

Rhamonic → harmonic

$$\underbrace{C(x(t))}_{\text{time-domain signal}} = \underbrace{F^{-1}\Big[\overbrace{\log\big(\overbrace{\underbrace{F(x(t))}_{\text{DFT}}}^{\text{spectrum}}\big)}^{\text{log spectrum.}}\Big]}_{\text{Cepstrum}}$$

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

Formalising speech

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

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

$$\underbrace{\log(X(t))}_{\text{Speech}} = \underbrace{\log(E(t))}_{\text{Glottal}} + \underbrace{\log(H(t))}_{\text{Vocal tract frequency Response}}$$

~~Gott~~

$$\Downarrow \text{IDFT}$$

$$X(t) = \underbrace{\cancel{E(t)}}_{\text{not interested}} + H(t)$$

(low pass lifter)

M-L approach → traditional
→ Deep learning.

Deep learning
↓
unstructured data.

Types of intelligent audio systems.

Traditional ML
↓
feature engineering

Deep learn
↓
automatic feature extractions

Extract audio features.

Time-domain feature pipeline
↓
Analogue sound
↓ ADC
digital signal.
↓
Framing (making frames of samples)
↑
Feature computation
↑
aggregation
↑
Feature value / vector

Typical values → (256 – ~~[illegible]~~ 8192)
powers of 2

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

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.

↑ frame size ↑ freq resolution ↓ time resolution

Duratⁿ of
1 ~~set~~ sample << human resolutⁿ

Frequency Domain feature pipeline

Analogue
↓ ADC
Quantization
↓
(overlapping frames to prevent losing signal during windowing) Framing
↓
Fourier transform ⟶ Frequency domain ⟶ Windowing ⟶ feature computatⁿ ⟶ aggregation ⟶ feature value / vector matrix

Frequency domain
↓
Spectral leakage.

- 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 → To make it periodic.
- Generates a periodic signal. → minimizes spectral leakage.

popular functⁿ: Hann window.

hop length: over lap betw ~~fr~~ ~~win~~ 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 segmentatn, music genre classification

Zero crossing Rate

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

  Speech Recognition

Fourier

- Compa
- for eac
- high
  a s

- Apply

Output

- DFT
  - Spe
  - N
- STFT
  - Sp
  - Co

Time /

↑

frame