Data Mining – UNDERSTANDING DATASET

djanloo – 23 set 22

Problem → Infer some features of the speaker given an audio sample

Dataset structure

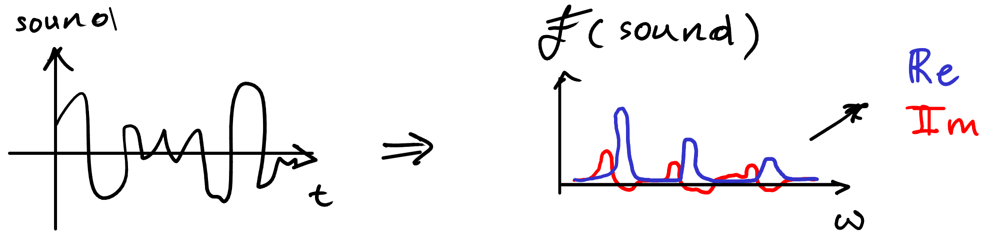
For each record the following macroscopic features are given :

- DUMB STUFF (M/F, EMOTION, etc)
- MEL-FREQ CEPSTRAL COEFFICIENTS (MFCC)
- SPECTRAL CENTROID (SC)
- SHORT TIME FOURIER TRANSFORM CHROMAGRAM (STFTC)

INTRO

MFCC EXPLAINED

⇒ aim at characterizing the human voice through periodic structures of the sound spectrum (CEPSTRUM)

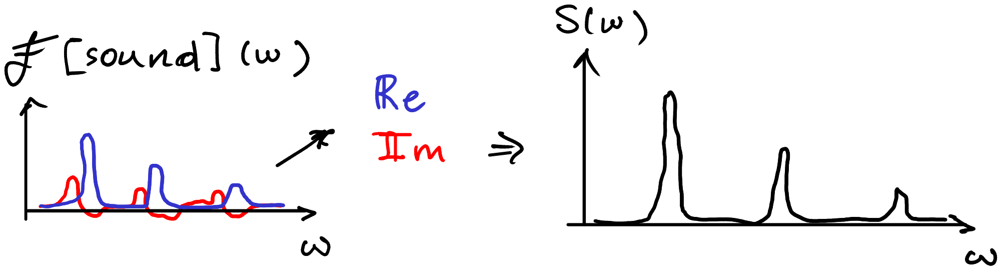


But the fourier transform is complex-valued
⇒ harder inference

Moreover the timbre of the sound has to do with the amplitudes of the upper harmonic range (overtones) that reflect in the periodic structure (ω_0 , $2\omega_0$, $3\omega_0$ ...) of the spectrum

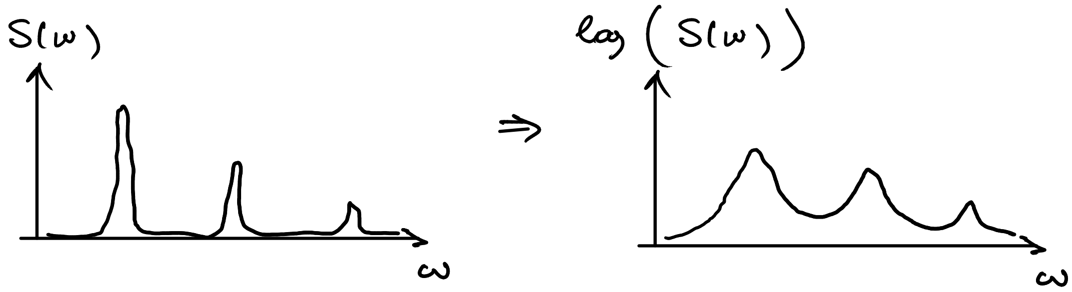
Thus the $\mathcal{F}$(sound) is taken in complex abs:

$$S(\omega) = \left| \mathcal{F}[\text{sound}](\omega) \right|^2$$ (power spectrum)

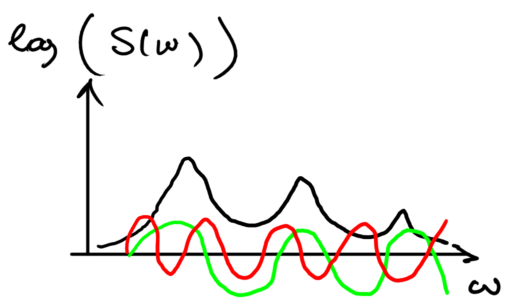


to get rid of the complex problem.

Furthermore, since the power spectrum has high peaks, the whole $S(\omega)$ gets log-rescaled

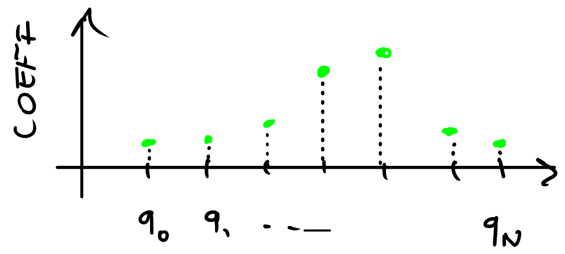


Since $\log(S(\omega))$ is a Real fuction, it can be inverse-transformed using DCT^{-1} (inverse discrete cosine transform, it's not clear WHICH DCT^{-1} is performed, but nevermind it's not so relevant) ⇒ extract $\underbrace{\text{"frequency of frequencies"}}_{\text{QUEFRENCY}}$



MATCHING QUEFRENCY → BIG COEFFICIENT
NOT MATCHING QUEFRENCY → SMALL COEFFIC.

This is done for a bunch of quefrencies $q_0, q_1, \dots, q_N\}$ obtaining the coeff array, which will be approximately monomodal:



$$\vec{c} = (c_1, c_2, c_3, \dots, c_N)$$

MFCC

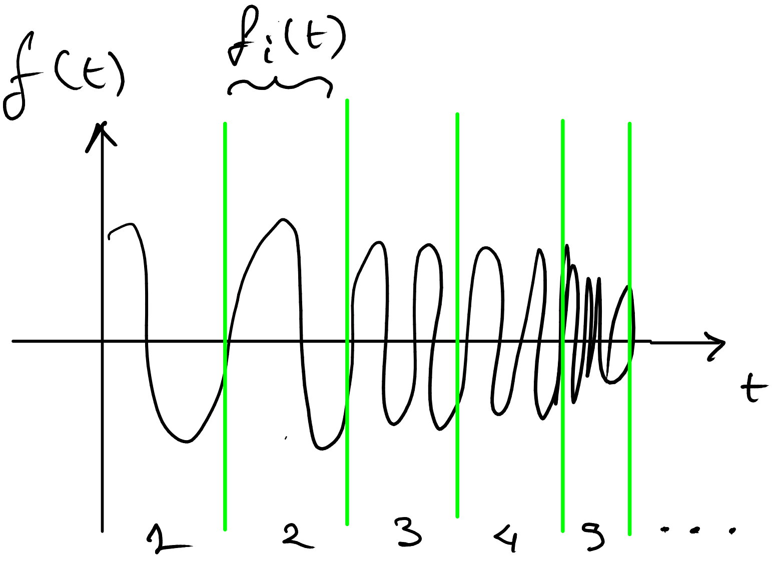
STUFF GIVEN IN THE DATASET: $\langle \vec{c} \rangle$ $\sigma(\vec{c})$

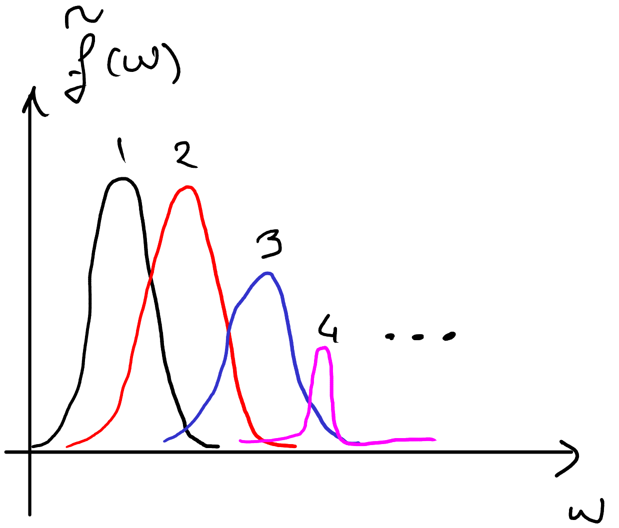
$\min(\vec{c})$ $\max(\vec{c})$

$\sqrt{\text{MFCC}}$

SPECTRAL CENTROID

signal divided in frames $\Rightarrow$ each frame gets fourier transformed

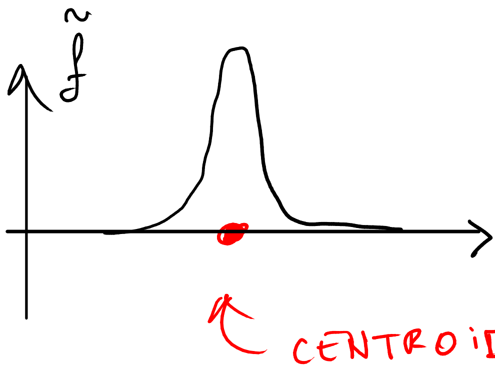


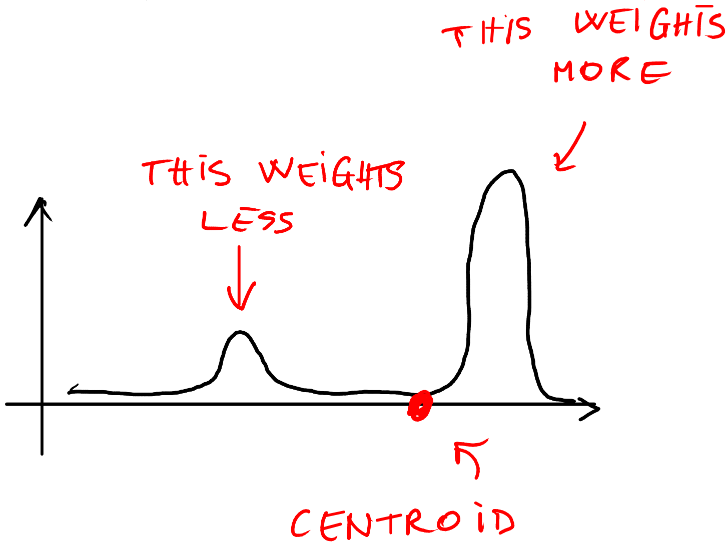


Each spectrum has a spectral centroid, that is the "center of mass" of the spectrum

$$\bar{\omega}_i \propto \int \omega \, \tilde{f}_i(\omega) \, d\omega$$

EXAMPLE





$\sqrt{\text{SC}}$

So for each frame a $\overline{\omega}_i$ is computed

STUFF GIVEN IN THE DATASET:

$$\langle \overline{\omega} \rangle \qquad \sigma_\omega \qquad \mathrm{KURT}(\omega) \qquad \mathrm{SKEW}(\omega)$$

$$\mathrm{KURT}(\omega) \to \propto \mu_4 \qquad \mathrm{SKEW}(\omega) \to \propto \mu_3$$

$\lceil$SC

STFTC

https://www.audiolabs-erlangen.de/resources/MIR/FMP/C3/C3S1_SpecLogFreq-Chromagram.html

Humans perceive notes that differ by an octave as similar.

Pitch is an arbitrary scale of perception:

$$p = 69 + \log_2 (f / f_0)$$

where $f_0 = 440\ Hz$ is the A4 note.

This definition of the pitch grants that two notes that differ by 12 pitch units have double freqs and vice-versa.

$\lceil$STFTC

FREQUENCY POOLING

freqs are continuous but notes are discrete $\Rightarrow$ "equal" interval pooling:

EXAMPLE $254\ Hz < f < 269\ Hz \Rightarrow C4$

$269\ Hz < f < 285\ Hz \Rightarrow C4^{\#}$

ETC

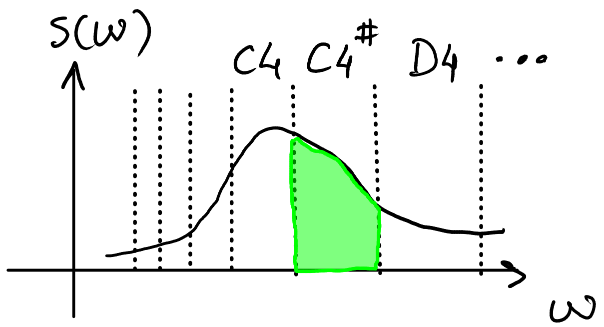
Frequency intervals are not equal but pitch intervals have unitary length:

$$59.5 < \text{pitch} < 60.5 \Rightarrow C4$$

$$60.5 < \text{pitch} < 61.5 \Rightarrow C4^{\#}$$

ETC

The sound intensity for each pitch is the integral over the freq. interval of the power spectrum $S(\omega)$



intensity of the $C4^{\#}$ note

STFTC

CHROMAGRAM

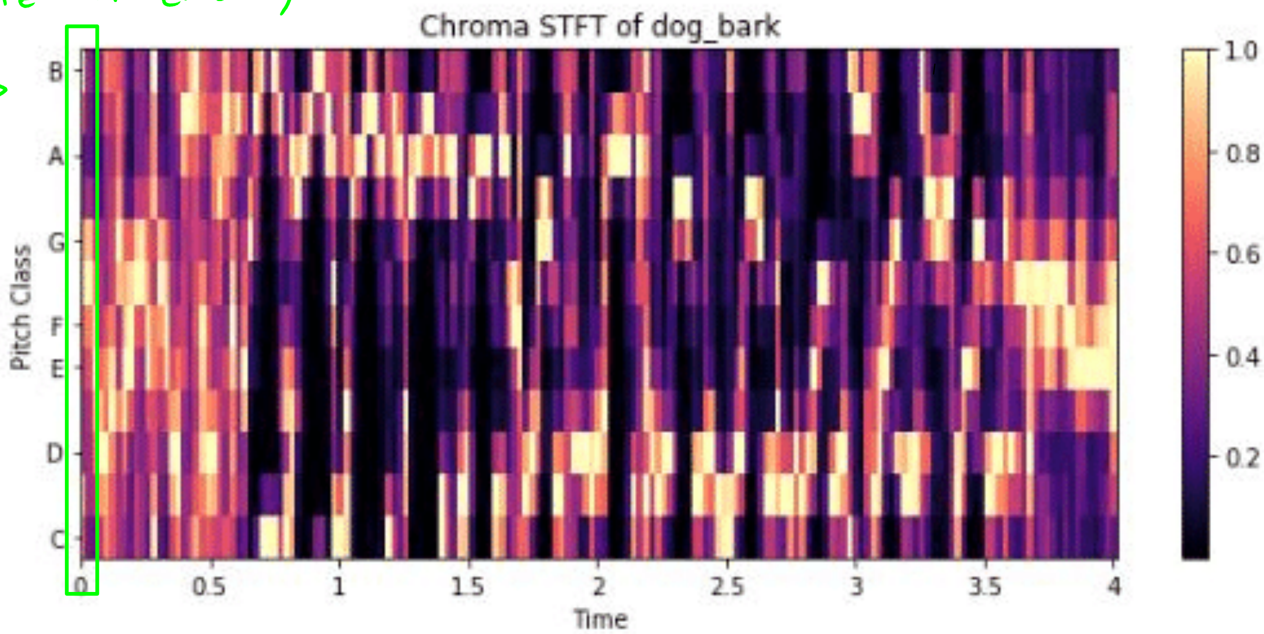
Since each note has an intensity and notes that differ by an octave have the same "color", it is possible to define a color intensity by:

$$I_C = \dots + I_{C4} + I_{C5} + I_{C6} + \dots$$

The chromagram displays this

This is done for all the small time intervals in which the sample is splitted.

NOTE INTENSITY IN THE FIRST TIME INTERVAL



STFTC