Автоматическая обработка звуковых образов.

Евгений Борисов

Обработка звуковых образов

Задачи

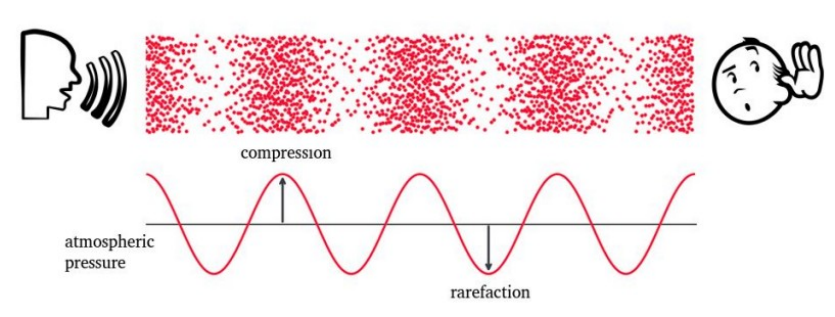
- Audio classification
- Speech recognition / speaker verification / speaker diarization
- Audio denoising / audio upsampling
- Music Information Retrieval (Mood Classification)
- Audio styling
- Audio synthesis

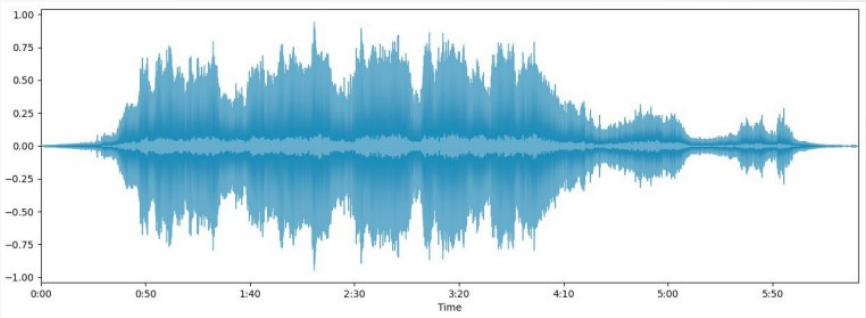
Обработка звуковых образов

звуковые волны

регулярные изменения давления воздуха

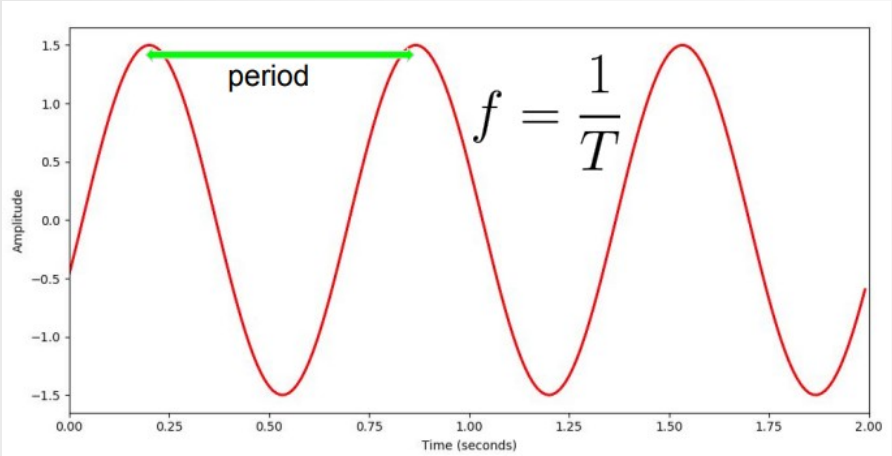
порождаются механической вибрацией объектов



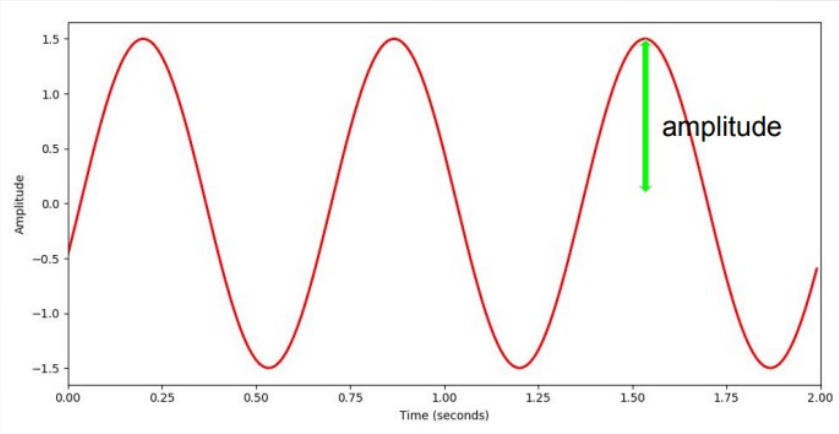


Обработка звуковых образов

частота



амплитуда



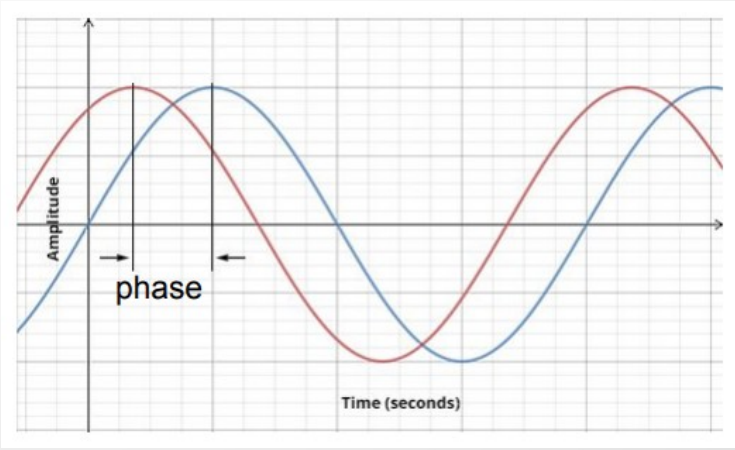
$$y(t) = A \cdot \sin(2\pi f t + \phi)$$

ϕ – амплитуда;
f – частота;
t – время;
ϕ – фаза;

Частота определяет высоту звука,

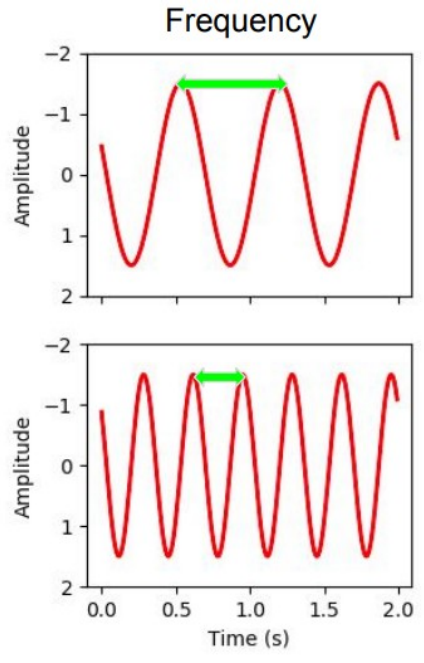
Амплитуда - громкость

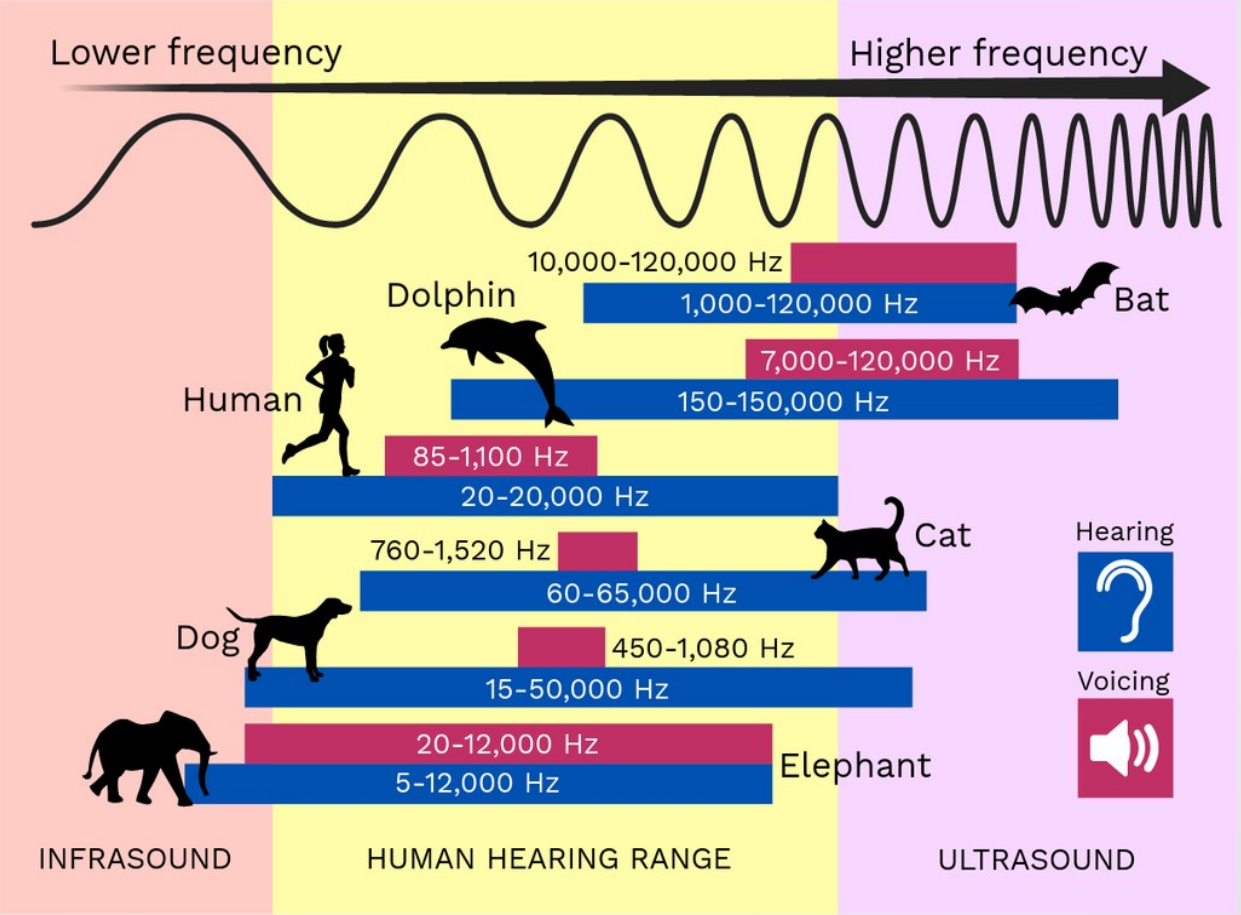
фаза



Обработка звуковых образов

высокие и низкие звуки





Обработка звуковых образов

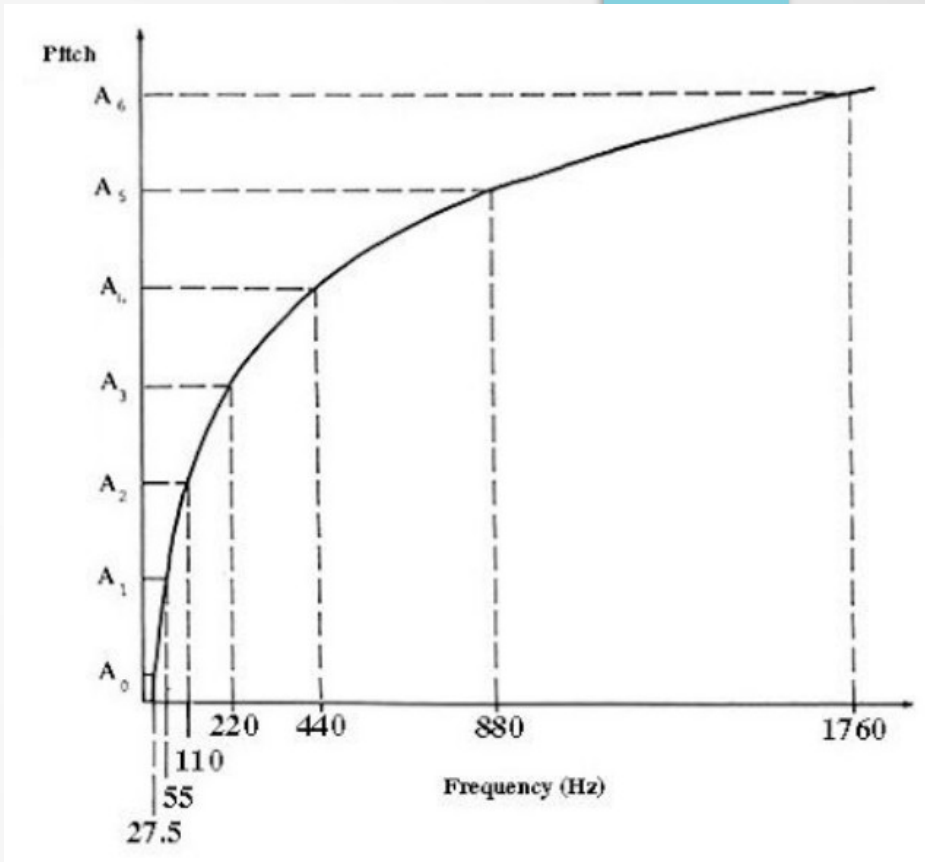
тон и частота

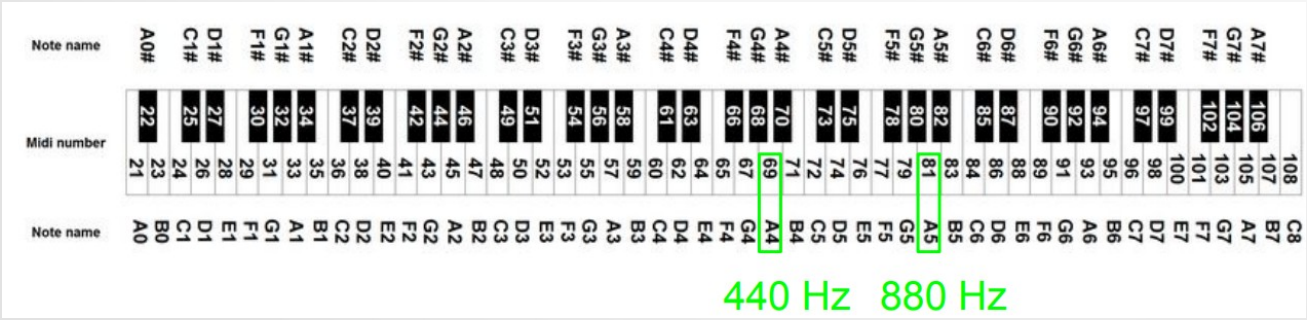
Тон - восприятие частоты человеком

логарифмическое восприятие,
человек различает низкие звуки лучше чем высокие

$$F(p) = 440 \cdot 2^{\frac{p-69}{12}}$$

Тембр — многомерная характеристика составляющих звука





Обработка звуковых образов

Громкость — субъективное восприятие человеком, зависит от возраста, частоты и др.

мощность звука - энергия передаваемая от источника в единицу времени (Вт, W)

интенсивность звука - мощность переданная на единицу площади (Вт/м2, W/m2)

уровень интенсивности — логарифмическое восприятие интенсивности человеком (dB)

порог слышимости , threshold of hearing (TOH)

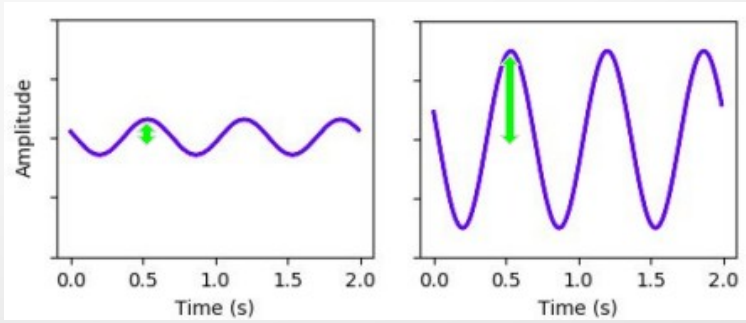
$$TOH = 10^{-12}\, W/m^2$$

болевой порог громкости , threshold of pain (TOP)

$$TOP = 10\, W/m^2$$

$$dB(I) = 10 \cdot \log_{10} \frac{I}{TOH}$$

$$dB(TOH) = 10 \cdot \log_{10} \frac{TOH}{TOH} = 10 \cdot \log_{10} 1 = 0$$

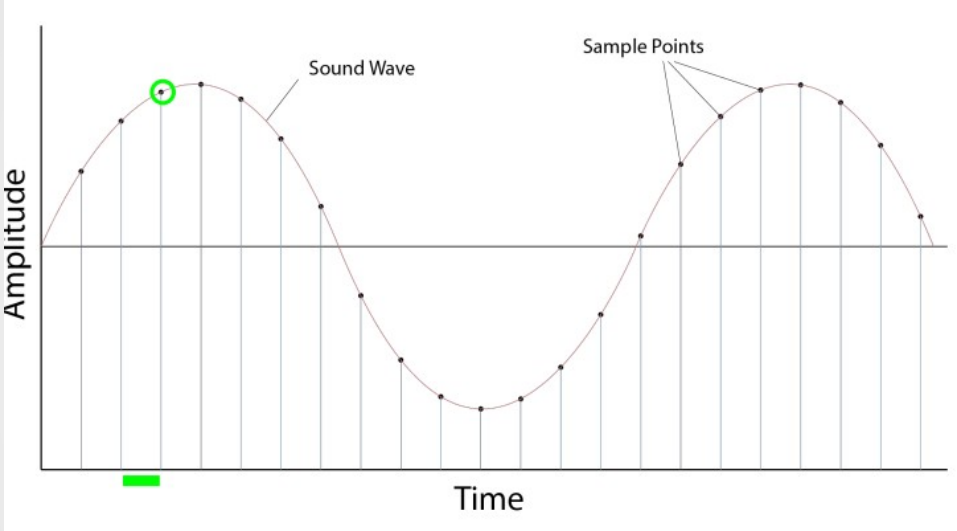


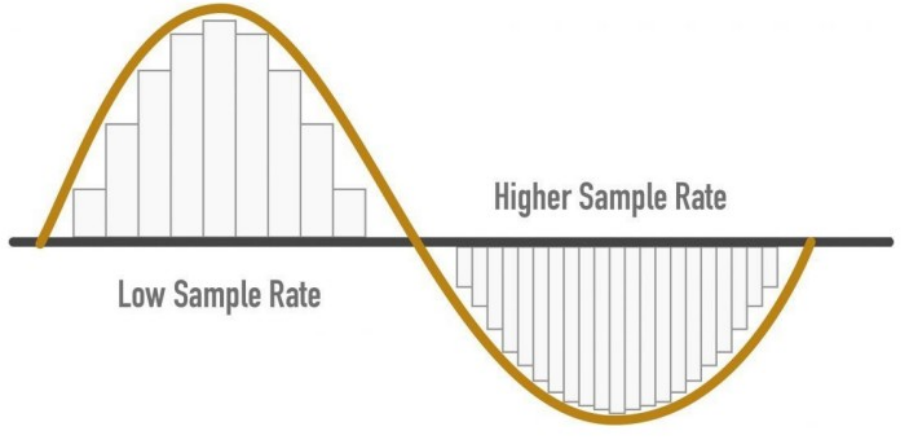
Обработка звуковых образов

представление звукового сигнала

Обработка звуковых образов

представление звукового сигнала — частота дискретизации





Sampling rate

$$Sr = \frac{1}{T}$$

Теорема Найквиста – Шенона – Котельникова

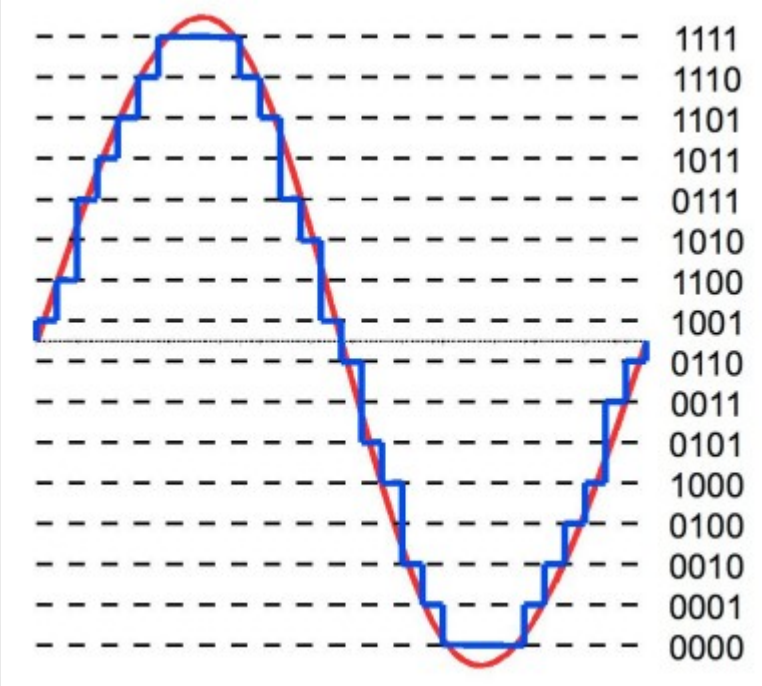
$$f_N = \frac{Sr}{2}$$

Если максимальная частота воспринимаемая человеком 22050Гц,

То для воспроизведения без (существенных) потерь минимальная частота дискретизации звукового сигнала (sample rate) должна быть 44100Гц (CD качество музыки)

Обработка звуковых образов

представление звукового сигнала - quantization, bit depth



bit depth - насколько точно будем кодировать уровень сигнала

CD качество

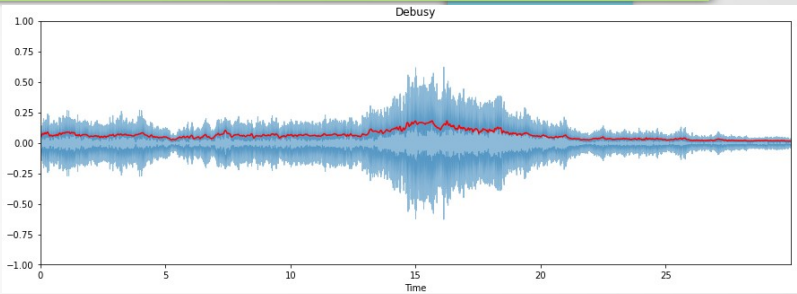
Sample rate = 44100 Hz
bit depth = 16 bits

$$2^{16} = 65536$$

Обработка звуковых образов

извлечение признаков из звуковых сигналов

- моментальная характеристика (точечная)
- характеристика сегмента (на отрезке)
- общая характеристика (всё целиком)



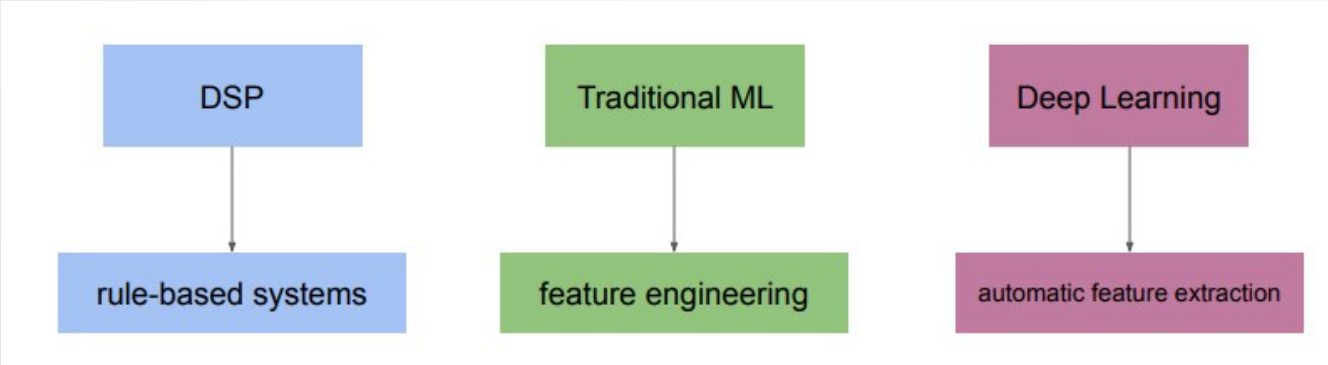
различные типы звуковых сигналов - разные наборы признаков

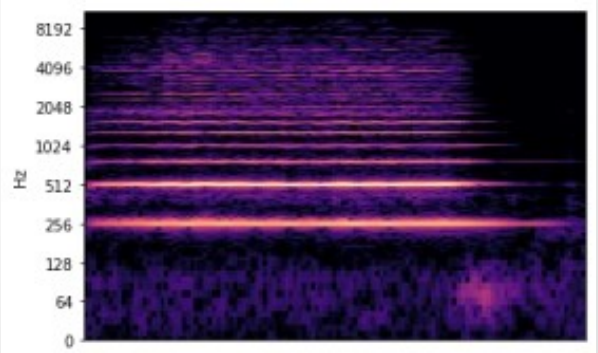
- Обработка музыкальных образов (Music Information Retrieval, MIR)

- Задача распознавания речи

- Прочее

- Amplitude envelope
- Root-mean square energy
- Zero crossing rate
- Spectrogram
- Spectral centroid
- Mel-spectrogram





Обработка звуковых образов

Типы аудиопризнаков:

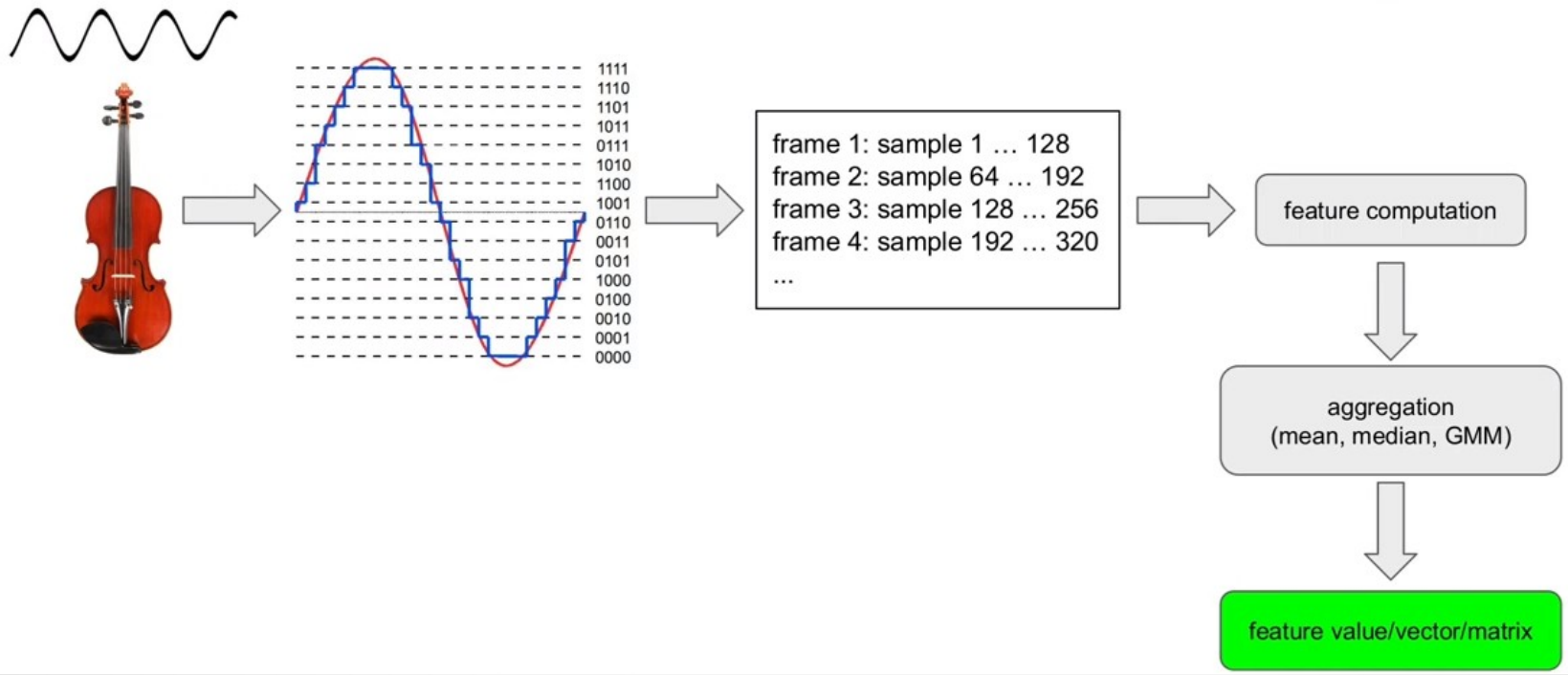
Временные характеристики (Time-domain features)

Частотные характеристики (frequency domain features)

Обработка звуковых образов

Временные характеристики (Time-domain features)

считаем характеристику для каждого фрейма и агрегируем результаты



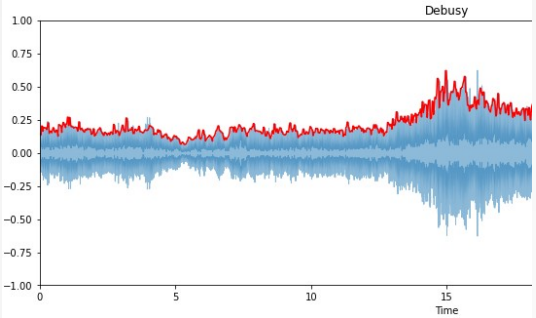
Обработка звуковых образов

Временные характеристики (Time-domain features)

считаем характеристику для каждого фрейма и агрегируем результаты

$Amplitude\ envelope\ (AE)$

$AE(t) = max_{k=t\cdot K}^{(t+1)\cdot K-1} s(k)$

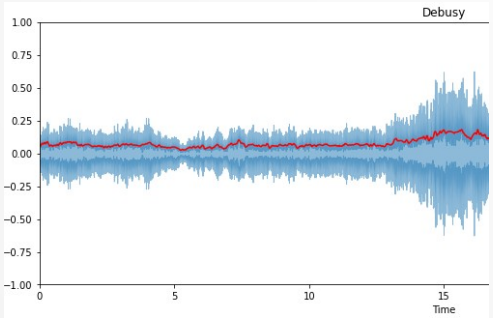


характеристика чувствительна к выбросам

можно использовать для определение жанра музыки

$Root-mean-square\ energy\ (RMS)$

$RMS(t) = \sqrt{\frac{1}{K}\sum_{k=t\cdot K}^{(t+1)\cdot K-1} s^2(k)}$



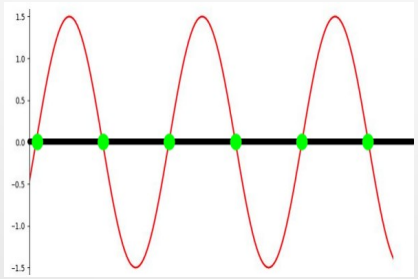
определяет уровень громкости

менее чувствительна к выбросам чем AE

можно использовать для сегментации аудио

$Zero-crossing\ rate\ (ZCR)$

$ZCR(t) = \frac{1}{2}\sum_{k=t\cdot K}^{(t+1)\cdot K-1} |sign(s(k)) - sign(s(k+1))|$



детектор ритмичной музыки

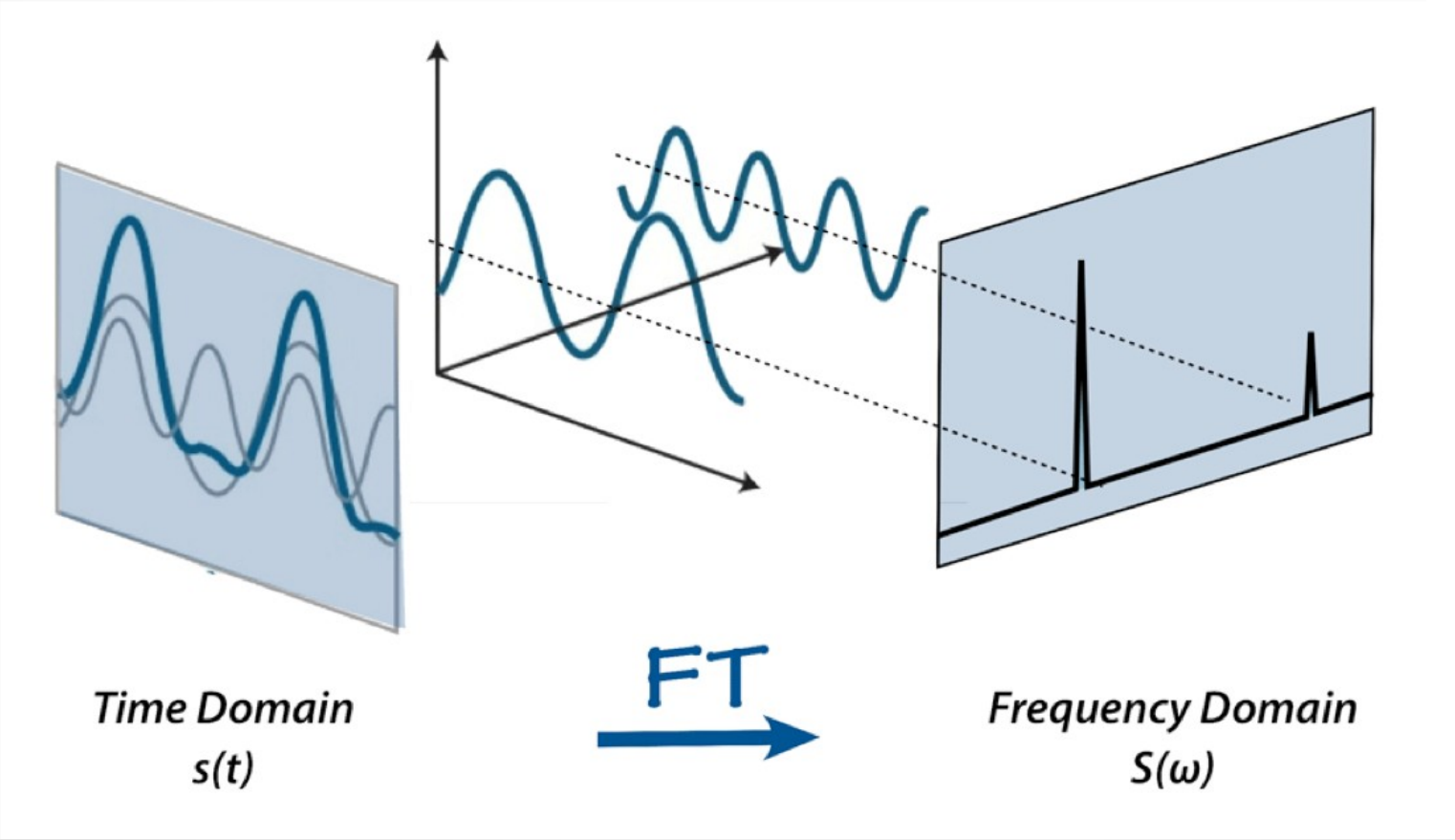
детектор голоса

Обработка звуковых образов

Частотные характеристики (frequency domain features)

преобразование Фурье - разбираем сигнал на частотные составляющие

$$G(f) = \int_{-\infty}^{\infty} g(t) e^{-2\pi i f t} dt.$$

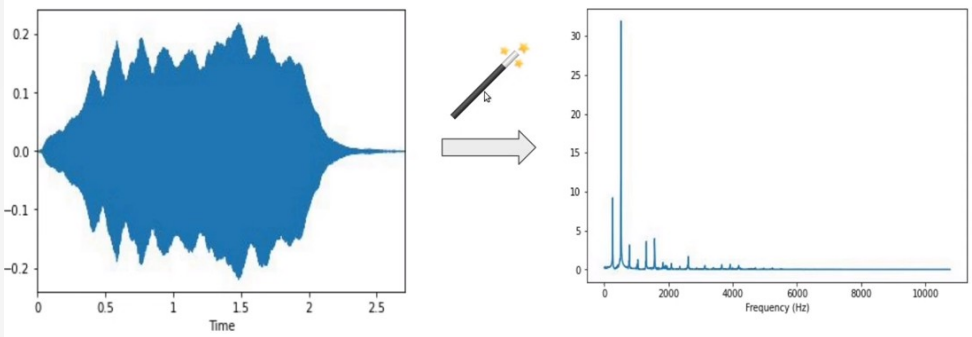


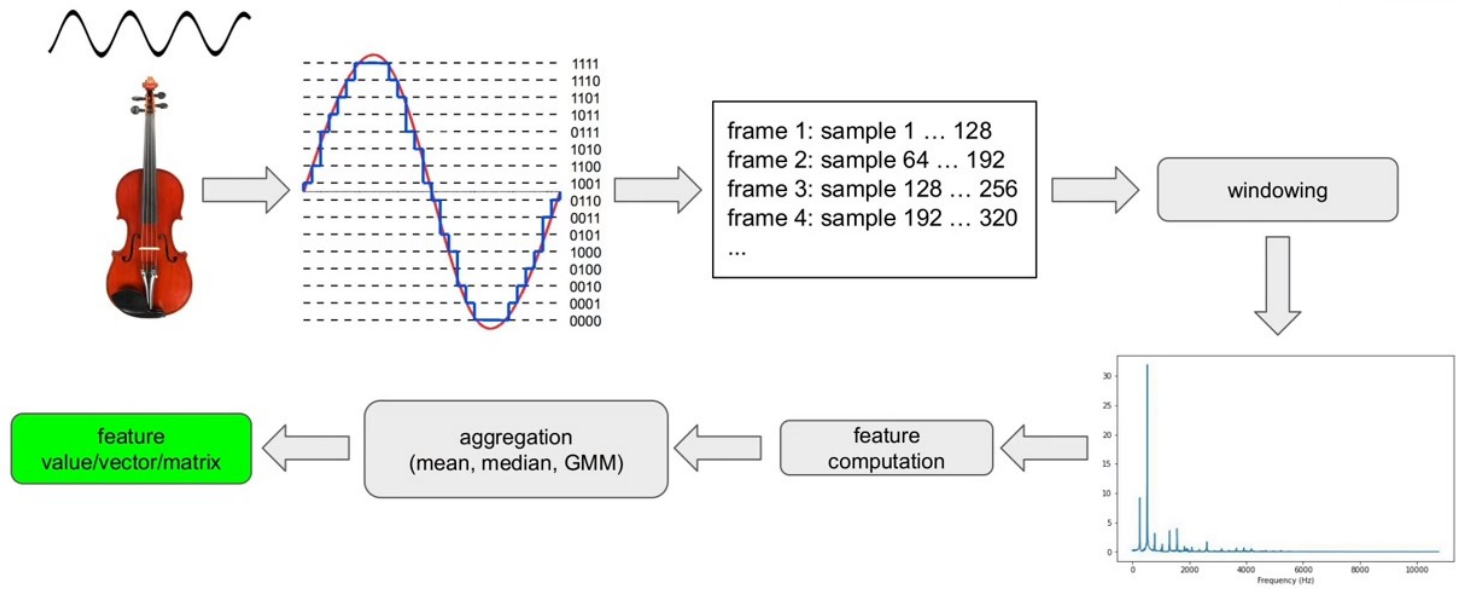
https://proglib.io/p/fourier-transform

Обработка звуковых образов

Частотные характеристики (frequency domain features)

разбираем сигнал на частотные составляющие, считаем характеристику и агрегируем результаты





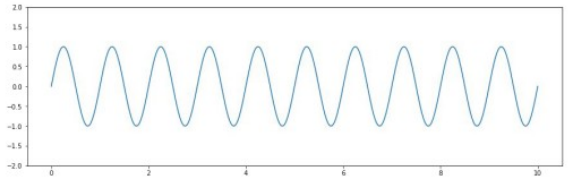
Обработка звуковых образов

Частотные характеристики (frequency domain features)

преобразование Фурье - разбираем сигнал на частотные составляющие

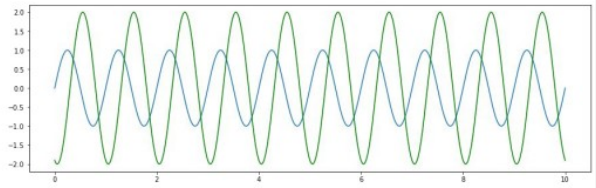
Pure tone of frequency *f*

$$g(t) = \int c_f \cdot e^{i2\pi ft} df$$



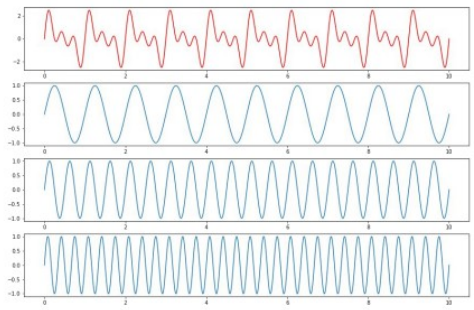
Weight pure tone with magnitude and add phase

$$g(t) = \int c_f \cdot e^{i2\pi ft} df$$



Add up all (weighted) sinusoids

$$g(t) = \int c_f \cdot e^{i2\pi ft} df$$



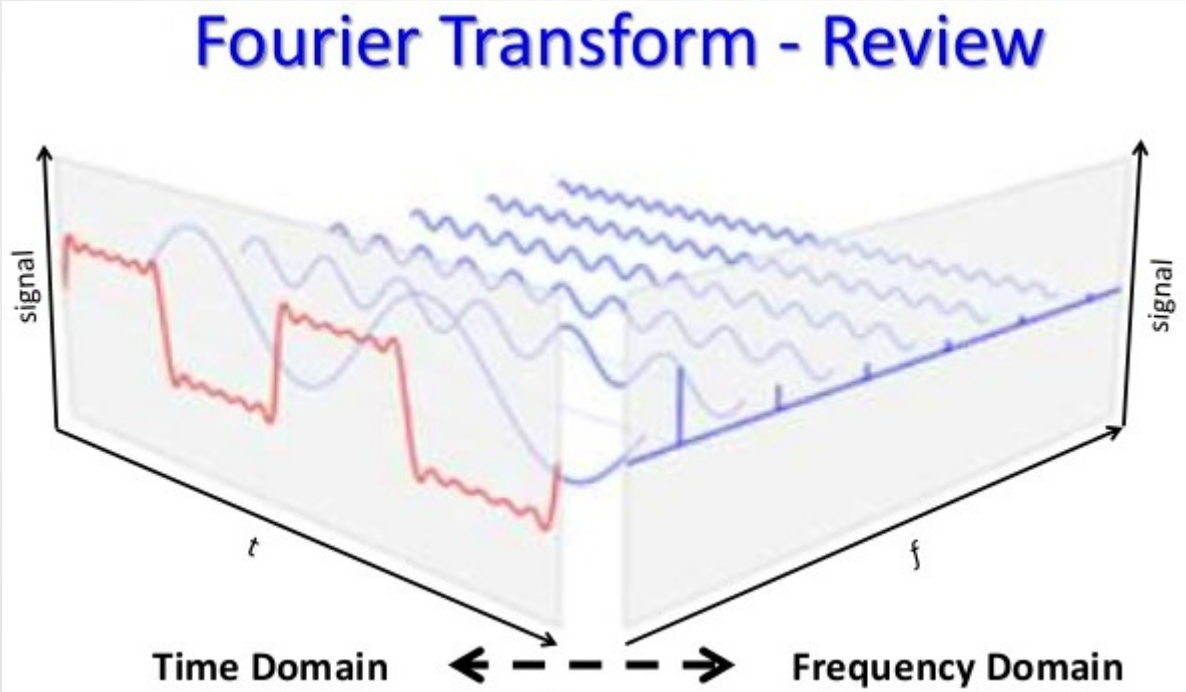
Обработка звуковых образов

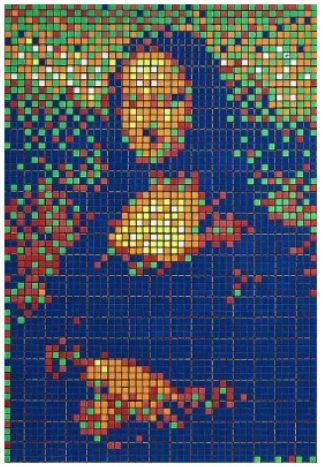
Частотные характеристики (frequency domain features)

Прямое и обратное преобразование Фурье

$$\hat{g}(f) = \int g(t) \cdot e^{-i2\pi ft} dt$$

$$g(t) = \int c_f \cdot e^{i2\pi ft} df$$

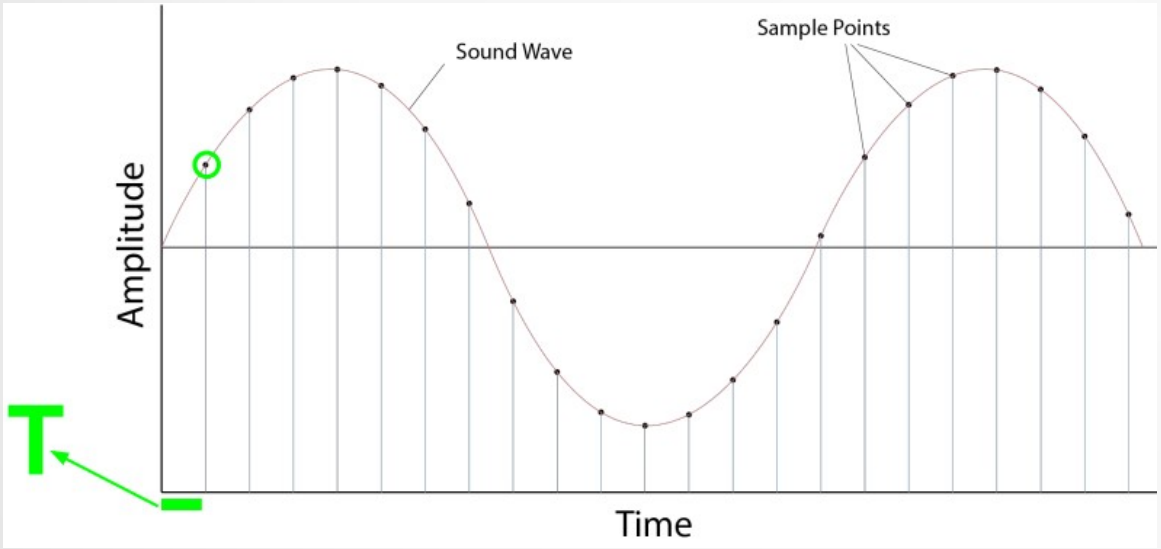


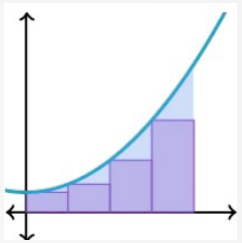


Обработка звуковых образов

Частотные характеристики (frequency domain features)

дискретное преобразование Фурье DFT





$$\hat{g}(f) = \int g(t) \cdot e^{-i2\pi ft} dt$$

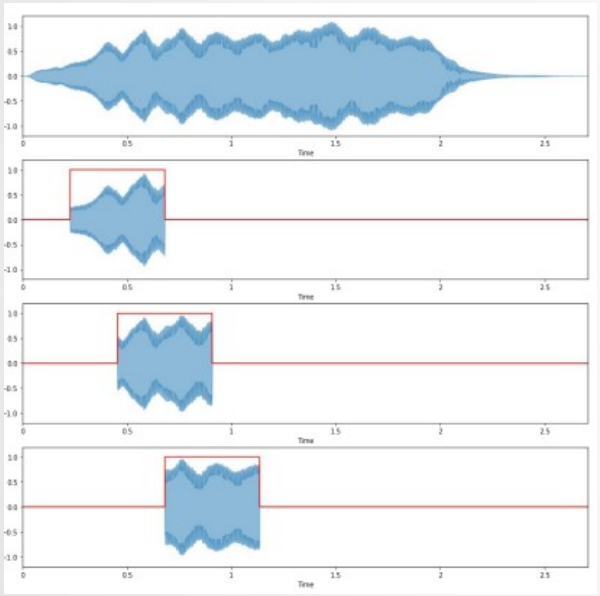
$$\hat{x}(f) = \sum_{n} x(n) \cdot e^{-i2\pi fn}$$

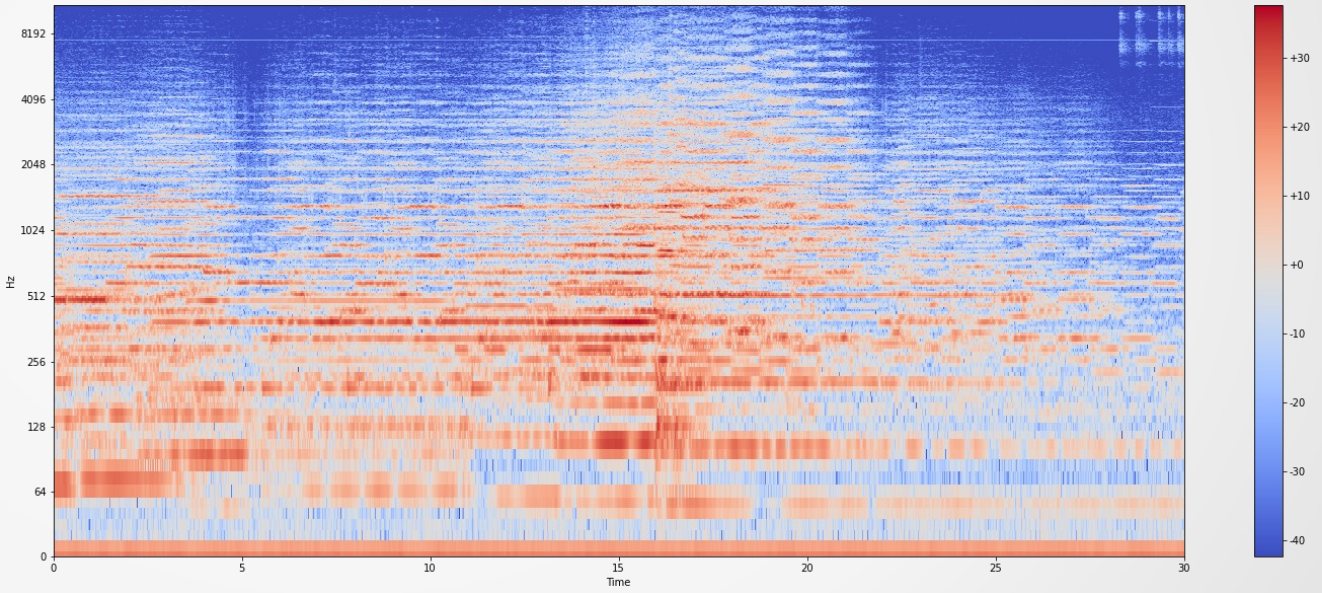
Обработка звуковых образов

Частотные характеристики (frequency domain features)

разбираем сигнал на частотные составляющие, считаем характеристику и агрегируем результаты

Спектрограммы — разбиваем образ на несколько перекрывающихся окон, в каждом применяем FT





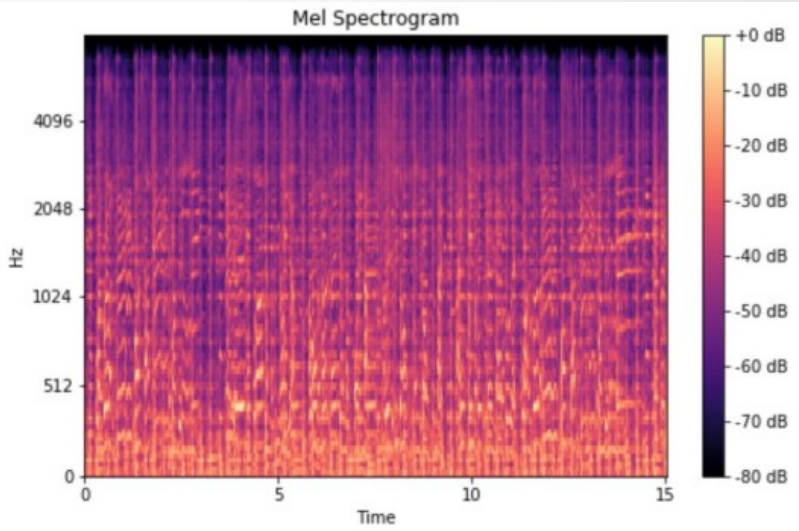
Обработка звуковых образов

Частотные характеристики (frequency domain features)

Мэл-спектрограммы

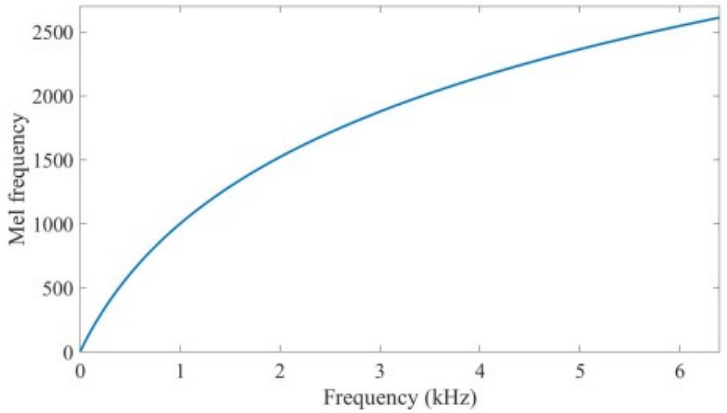
логарифмическое восприятие,
человек различает низкие частоты лучше чем высокие

выполняем логарифмическое преобразование Hz → Mel



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

$$f = 700(10^{m/2595} - 1)$$

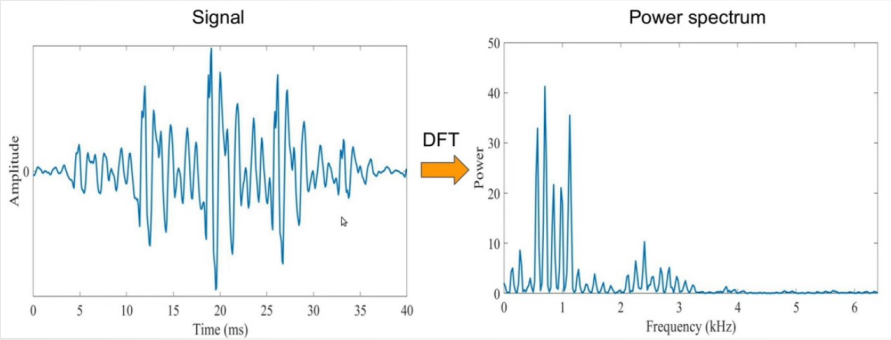


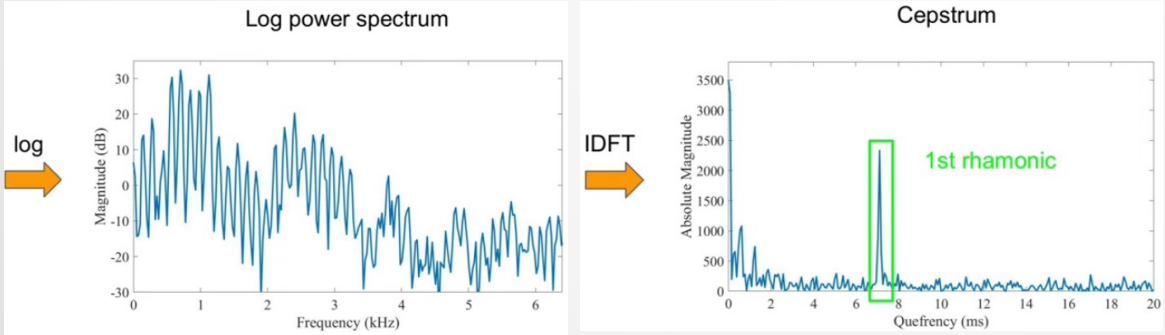
Обработка звуковых образов

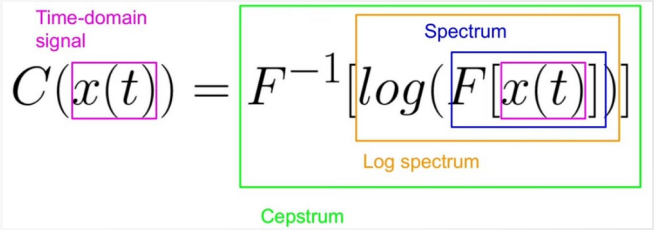
Частотные характеристики (frequency domain features)

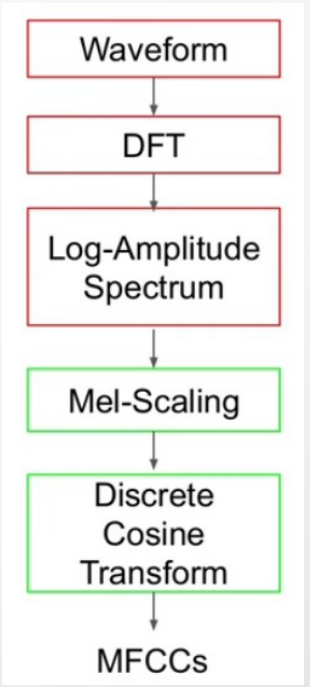
Мэл-кепстральные коофициенты (Mel-Frequency Cepstral Coefficients, MFCC)

Spectrum - **Ceps**trum









Mel-Frequency Cepstral Coefficients Explained Easily
https://www.youtube.com/watch?v=4_SH2nfbQZ8

Обработка звуковых образов

Обработка музыкальных образов
(Music Information Retrieval, MIR)

задачи
- разделение звучания музыкальных инструментов
- классификация музыкальных образов
- оценка настроения музыкального образа

характеристики

- ритм
- тон
- тембр
- гармоничность

Обработка звуковых образов

Литература

git clone https://github.com/mechanoid5/ml_lectorium

Дауни Аллен Цифровая обработка сигналов на языке Python. 2017
Allen B.Downey Think DSP. Digital Signal Processing in Python. 2014

Meinard Müller, Brian McFee Instructional material for the Music Information Retrieval Workshop at CCRMA, Stanford University, 2014-18.
https://github.com/bmcfee/stanford-mir

Valerio Velardo The Sound of AI : Audio Signal Processing for Machine Learning
https://www.youtube.com/playlist?list=PL-wATfeyAMNqIee7cH3q1bh4QJFAaeNv0

https://github.com/musikalkemist/AudioSignalProcessingForML