Introduction to Digital Signal Processing

NATIONAL RESEARCH UNIVERSITY

2022
Deep Learning in Audio Processing

Daniil Ivanov

Sound representation

What is sound and how to store it in memory?

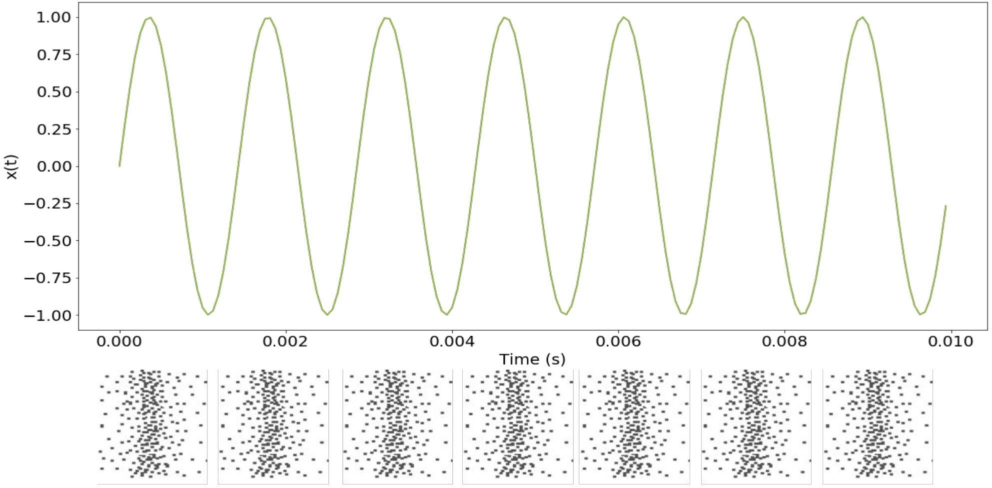
- **Sound representation**
- Motivation for spectrograms
- Fourier Transform
- Discrete Fourier Transform
- Short Time Fourier Transform
- Spectrogram
- Mel scale
- MFCC

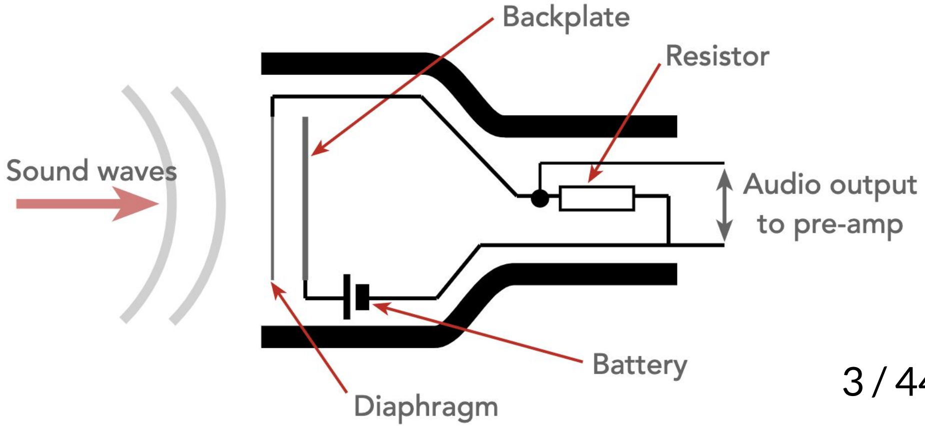
What is sound?

- **Sound wave** is the pattern of **oscillations** caused by the movement of energy traveling through the air

- **Microphone** picks up these air **oscillations** and converts them into electrical vibrations

- These **oscillations** are converted into an **analog** signal and then a **digital** signal

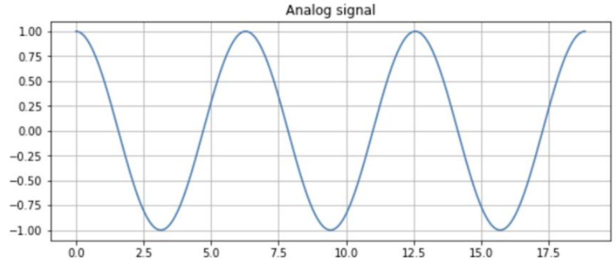


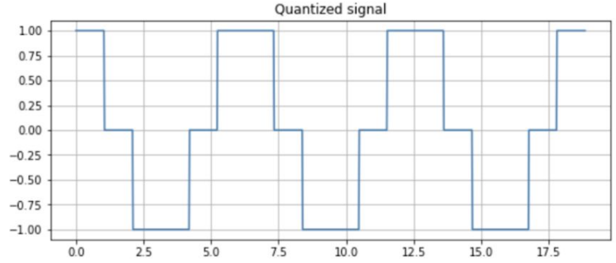


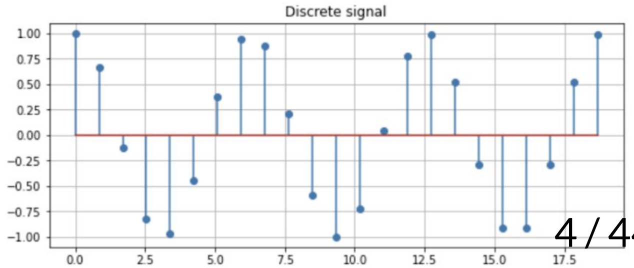
https://pudding.cool/2018/02/waveforms/
https://blog.accusonus.com/science-of-sound/how-microphones-work

How is sound stored in the computer?

- The **analog** signal is discretized, quantized and encoded
- An analog signal is **discretized** in that the signal is represented as a sequence of values taken at discrete points in time **t** with step **d**
- **Quantisation** of a signal consists in splitting the range of signal values into **N** levels in increments of **d** and selecting for each reference the level that corresponds to it
- Signal **encoding** is just a way of presenting the signal in a more compact form



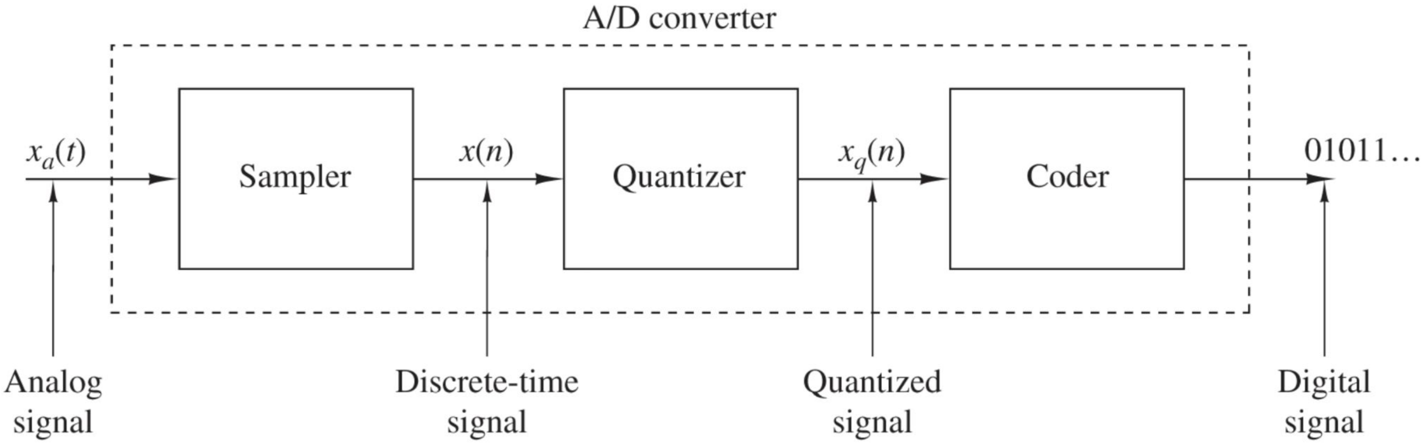




https://github.com/markovka17/dla/blob/master/week01/dsp.ipynb
https://web.sonoma.edu/esee/courses/ee442/archives/sp2019/lectures/lecture09_pcm.pdf

Analog-to-Digital Conversion

- Converting analog signals to a sequence of numbers having finite precision
- Corresponding devices are called A/D converters (ADCs)

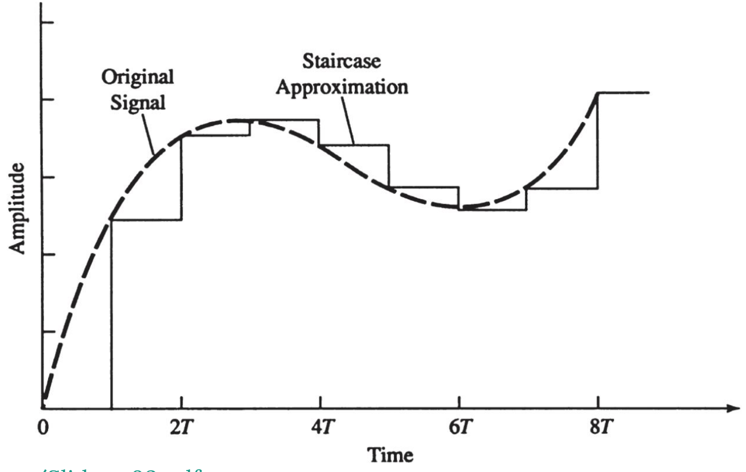


https://is.muni.cz/el/1433/jaro2012/PA190/um/Slides_02.pdf

Digital-to-Analog Conversion

- Process of converting a digital signal into an analog signal

- Interpolation
  - Connecting dots in a digital signal
  - Approximations: zero-order hold (staircase), linear, quadratic, and so on



https://is.muni.cz/el/1433/jaro2012/PA190/um/Slides_02.pdf

What other characteristics are there?

- **Sample rate (SR)** - number of audio samples per one second (e.g. 8 kHz, 22.05 kHz, 44.1 kHz)
- **Sample size** - number of bits per one sample (e.g. 8, 16, 25, 32 bits)
- **Number of channels** -- how many signals we record in parallel (e.g. mono(1), stereo(2))

8000 Hz

The international [G.711](#) standard for audio used in telephony uses a sample rate of 8000 Hz (8 kHz). This is enough for human speech to be comprehensible.

44100 Hz

The 44.1 kHz sample rate is used for compact disc (CD) audio. CDs provide uncompressed 16-bit stereo sound at 44.1 kHz. Computer audio also frequently uses this frequency by default.

48000 Hz

The audio on DVD is recorded at 48 kHz. This is also often used for computer audio.

96000 Hz

High-resolution audio.

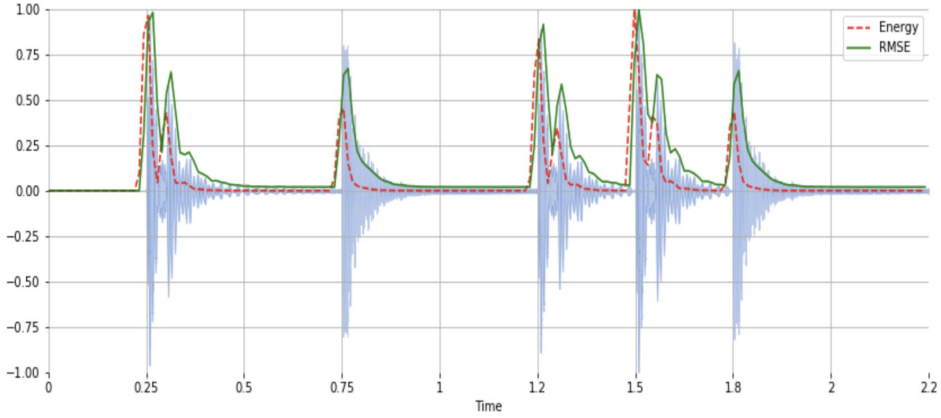
192000 Hz

Ultra-high resolution audio. Not commonly used yet, but this will change over time.

https://developer.mozilla.org/en-US/docs/Web/Media/Formats/Audio_concepts

What other characteristics are there?

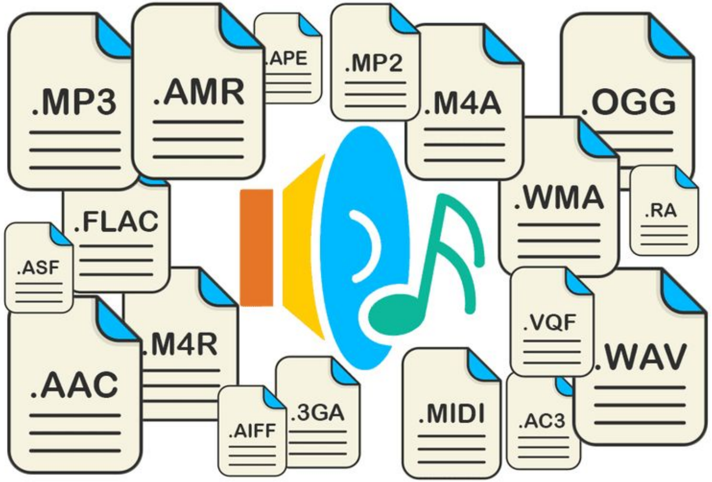
- Assume **f(n)** is our signal where **n** is time
- Power of signal is $f^2(n)$
- Energy of signal (**E**) is $\sum f^2(n)$
- In practice estimated by some window
- Energy in **decibels:** $10 \log_{10} E$
- $\mathrm{SNR}_{dB} = 10 \log_{10} \frac{E_{\text{signal}}}{E_{\text{noise}}}$



https://musicinformationretrieval.com/energy.html

What about audio formats?

- Non-compressed formats: **WAV, AIFF, etc.**
- Lossless compression(2:1) : **FLAC, ALAC, etc.**
- Lossy compression(10:1) : **MP3, Opus, etc**
- **Bit rate** measure a degree of compression. Number of bit that are conveyed or processed per **unit of time.**



https://en.wikipedia.org/wiki/Audio_file_format

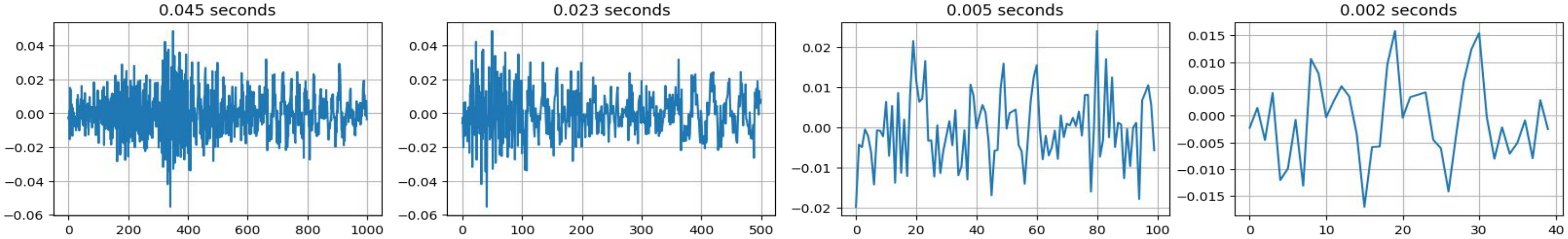
Motivation for spectrograms

Why not just use wave representation for ML?

- Sound representation
- **Motivation for spectrograms**
- Fourier Transform
- Discrete Fourier Transform
- Short Time Fourier Transform
- Spectrogram
- Mel scale
- MFCC

Problems with the waveform

- One letter/sound consists of 2000-4000 amplitudes, so they are expensive to process and store



- No "invariant" regarding noise and transformations
- Periodical nature of audio signals

Voice

Voice + Noise

Complex waves as a sum of sigmoids

We want to represent a periodic function as a sum of sigmoids with different periods (frequencies), shifts and amplitudes.

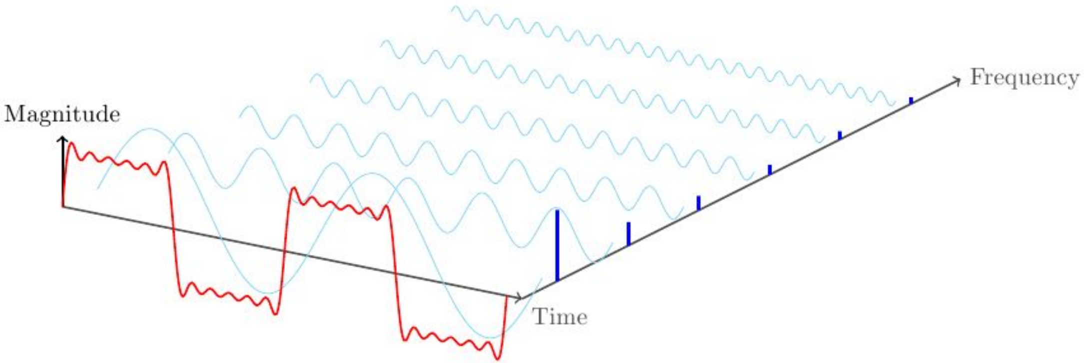
$$f(x) = A_1 * sin(freq_1 x + \phi_1) + \dots$$

$$\dots$$

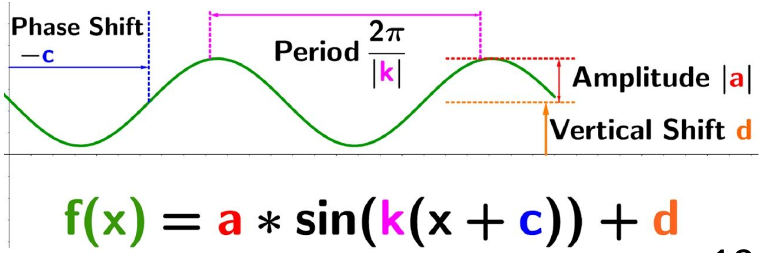
$$\dots + A_n * sin(freq_n x + \phi_n)$$

And for audio processing we are only interested in:

?????



Parameters of a sine wave



https://en.wikipedia.org/wiki/Sine_wave

Complex waves as a sum of sigmoids

We want to represent a periodic function as a sum of sigmoids with different periods (frequencies), shifts and amplitudes.

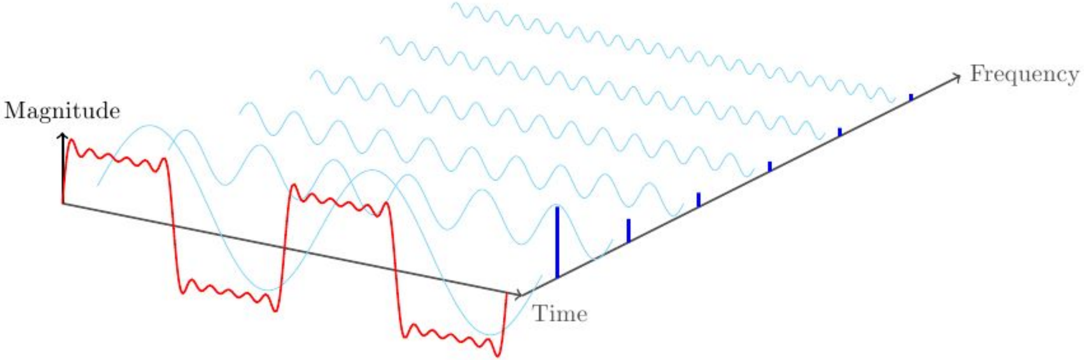
$$f(x) = A_1 * sin(freq_1 x + \phi_1) + ...$$

$$...$$

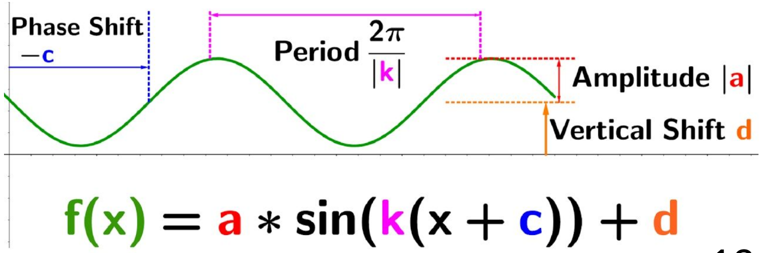
$$... + A_n * sin(freq_n x + \phi_n)$$

And for audio processing we are only interested in:

- **Frequencies**
- **Amplitudes**



Parameters of a sine wave



https://en.wikipedia.org/wiki/Sine_wave

Fourier Transform (FT)

How to factorize a periodic function into a sum of sine-waves?

- Sound representation
- Motivation for spectrograms
- **Fourier Transform**
- Discrete Fourier Transform
- Short Time Fourier Transform
- Spectrogram
- Mel scale
- MFCC

Fourier Transform

- The **Fourier transform(FT)** is a mathematical formula that allows us to decompose a signal into its individual **frequencies** and the frequency's **amplitude**

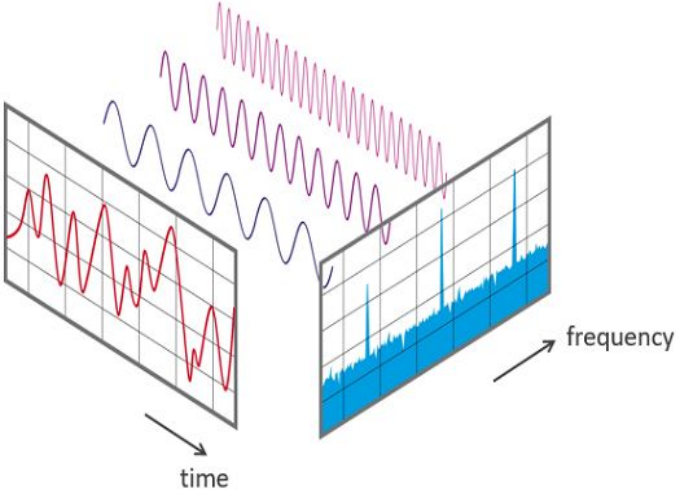
- FT transfer a signal from real-valued function of the **time domain** to a complex-valued function of **frequency domain**

Fourier transform integral

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, \quad \forall\, \xi \in \mathbb{R}.$$

$$f : \mathbb{R} \to \mathbb{R}$$

$$\hat{f} : \mathbb{R} \to \mathbb{C}$$



- The function must meet the following conditions:
  - to be **bounded**
  - to be **absolutely integrable**
  - to have a **finite number** of minimas, maximas and discontinuities

https://youtu.be/spUNpyF58BY

Fourier Transform

- The **Fourier transform(FT)** is a mathematical formula that allows us to decompose a signal into its individual **frequencies** and the frequency's **amplitude**

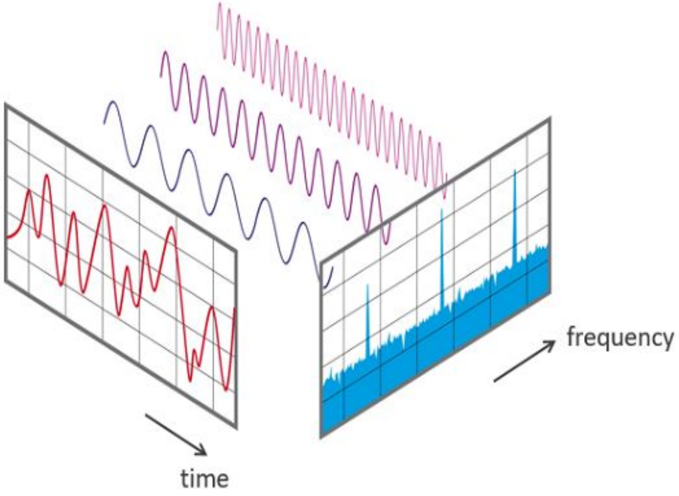
- FT transfer a signal from real-valued function of the **time domain** to a complex-valued function of **frequency domain**

- Fourier transform integral

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, \quad \forall\, \xi \in \mathbb{R}.$$

Frequency　Original signal

$$f : \mathbb{R} \to \mathbb{R}$$
$$\hat{f} : \mathbb{R} \to \mathbb{C}$$



https://youtu.be/spUNpyF58BY

Inverse Fourier Transform

Fourier transform integral

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, \quad \forall\, \xi \in \mathbb{R}.$$

Fourier inversion integral

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)\, e^{i2\pi\xi x}\, d\xi, \quad \forall\, x \in \mathbb{R},$$

https://en.wikipedia.org/wiki/Fourier_transform

Inverse Fourier Transform

Fourier transform integral

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, \quad \forall\, \xi \in \mathbb{R}.$$

Fourier inversion integral

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)\, e^{i2\pi\xi x}\, d\xi, \quad \forall\, x \in \mathbb{R},$$

$$= 2\int_{0}^{\infty} \mathrm{Re}\left(\hat{f}(\xi) \cdot e^{i2\pi\xi x}\right) d\xi$$

Property of FT

$$\hat{f}(\xi) = \begin{cases} \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, & \xi \geq 0 \\ \hat{f}^{*}(|\xi|) & \xi < 0, \end{cases}$$

https://en.wikipedia.org/wiki/Fourier_transform

Inverse Fourier Transform

Fourier transform integral

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, \quad \forall\, \xi \in \mathbb{R}.$$

Fourier inversion integral

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)\, e^{i2\pi\xi x}\, d\xi, \quad \forall\, x \in \mathbb{R},$$

$$= 2\int_{0}^{\infty} \mathrm{Re}\left(\hat{f}(\xi) \cdot e^{i2\pi\xi x}\right) d\xi$$

$$= 2\int_{0}^{\infty} \left(\mathrm{Re}(\hat{f}(\xi)) \cdot \cos(2\pi\xi x) - \mathrm{Im}(\hat{f}(\xi)) \cdot \sin(2\pi\xi x)\right) d\xi.$$

Property of FT

$$\hat{f}(\xi) = \begin{cases} \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi\xi x}\, dx, & \xi \geq 0 \\ \hat{f}^{*}(|\xi|) & \xi < 0, \end{cases}$$

Euler's formula

$$e^{jx} = \cos x + j \sin x$$

https://en.wikipedia.org/wiki/Fourier_transform

Inverse Fourier Transform

$$= 2\int_0^\infty \left(\mathrm{Re}(\hat{f}\,(\xi)) \cdot \cos(2\pi\xi x) - \mathrm{Im}(\hat{f}\,(\xi)) \cdot \sin(2\pi\xi x)\right) d\xi.$$

https://en.wikipedia.org/wiki/Fourier_transform

Inverse Fourier Transform

$$= 2\int_0^{\infty} \left(\mathrm{Re}(\hat{f}(\xi)) \cdot \cos(2\pi\xi x) - \mathrm{Im}(\hat{f}(\xi)) \cdot \sin(2\pi\xi x)\right) d\xi.$$

Trigonometry

$$A\ \cos(\omega t + \phi) = B\cos(\omega t) + C\sin(\omega t)$$
$$A = \sqrt{B^2 + C^2},\ \tan\phi = \frac{C}{B}$$

Example

$$\hat{f}(\xi_1) = c_1$$
$$\hat{f}(\xi_2) = c_2$$

$$f(x) = A_1\cos(2\pi\xi_1 x + \phi_1) + A_2\cos(2\pi\xi_2 x + \phi_2)$$
$$A_k = \sqrt{\mathrm{Re}(\mathrm{c_k})^2 + \mathrm{Im}(\mathrm{c_k})^2}$$
$$\tan\phi = \frac{C}{B}$$

Inverse Fourier Transform

$$= 2\int_0^{\infty} \left(\mathrm{Re}(\hat{f}(\xi)) \cdot \cos(2\pi\xi x) - \mathrm{Im}(\hat{f}(\xi)) \cdot \sin(2\pi\xi x)\right) d\xi.$$

Trigonometry

$$A\ \cos(\omega t + \phi) = B\cos(\omega t) + C\sin(\omega t)$$

$$A = \sqrt{B^2 + C^2},\ \tan\phi = \frac{C}{B}$$

Example

$$\hat{f}(\xi_1) = c_1$$

$$\hat{f}(\xi_2) = c_2$$

$$f(x) = A_1\cos(2\pi\xi_1 x + \phi_1) + A_2\cos(2\pi\xi_2 x + \phi_2)$$

$$A_k = \sqrt{\mathrm{Re}(c_k)^2 + \mathrm{Im}(c_k)^2}$$

$$\tan\phi = \frac{C}{B}$$

https://en.wikipedia.org/wiki/Fourier_transform

Discrete Fourier Transform (DFT)

How to calculate Fourier Transform in practice?

- Sound representation
- Motivation for spectrograms
- Fourier Transform
- **Discrete Fourier Transform**
- Short Time Fourier Transform
- Spectrogram
- Mel scale
- MFCC

Discrete Fourier transform

$$X = \mathbf{M}x$$

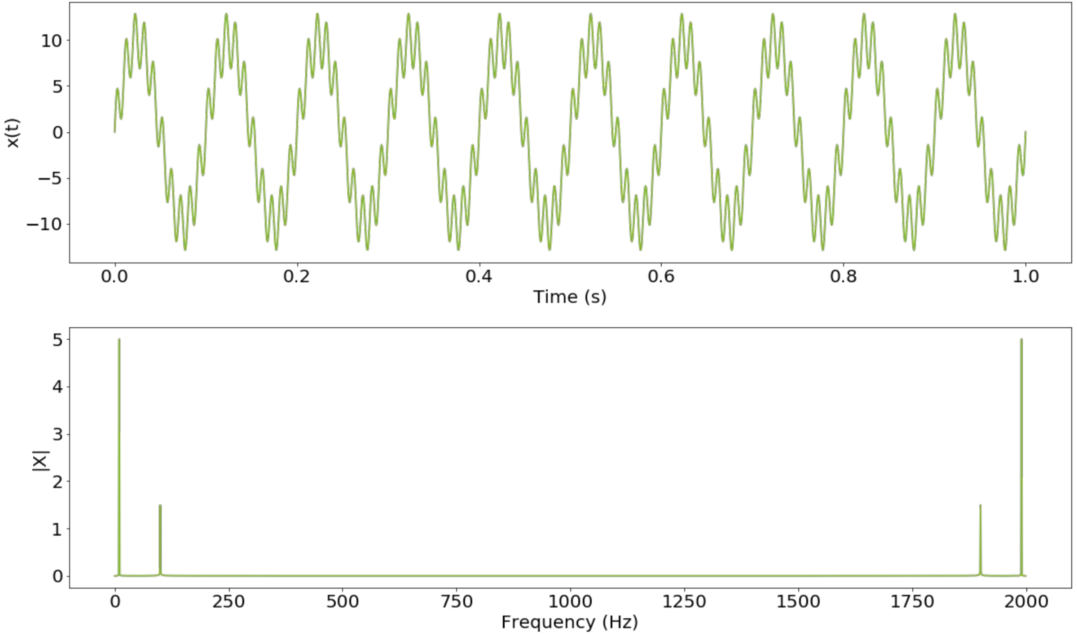
$$M_{mn} = \exp\left(-2\pi i \frac{(m-1)(n-1)}{N}\right)$$

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & e^{-\frac{2\pi i}{N}} & e^{-\frac{4\pi i}{N}} & e^{-\frac{6\pi i}{N}} & \dots & e^{-\frac{2\pi i}{N}(N-1)} \\ 1 & e^{-\frac{4\pi i}{N}} & e^{-\frac{8\pi i}{N}} & e^{-\frac{12\pi i}{N}} & \dots & e^{-\frac{2\pi i}{N}2(N-1)} \\ 1 & e^{-\frac{6\pi i}{N}} & e^{-\frac{12\pi i}{N}} & e^{-\frac{18\pi i}{N}} & \dots & e^{-\frac{2\pi i}{N}3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{-\frac{2\pi i}{N}(N-1)} & e^{-\frac{2\pi i}{N}2(N-1)} & e^{-\frac{2\pi i}{N}3(N-1)} & \dots & e^{-\frac{2\pi i}{N}(N-1)^2} \end{pmatrix}$$

Example of DFT

$$F = 2kHz$$

$$f(t) = 10\sin(2\pi 10t) + 3\sin(2\pi 100t)$$

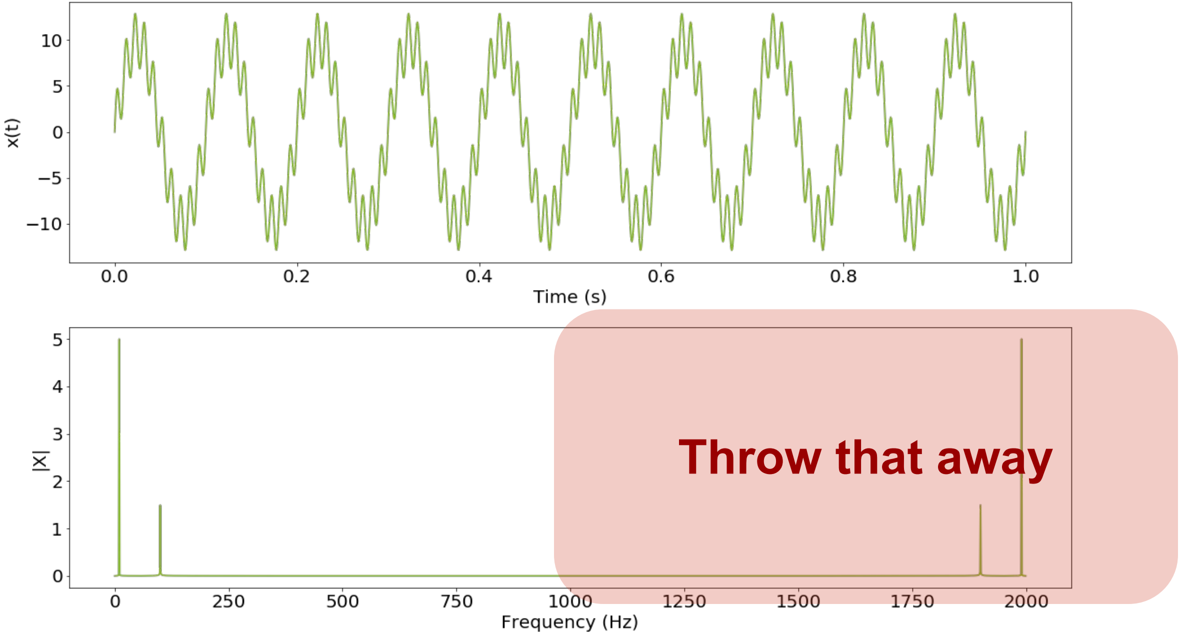


https://www.gaussianwaves.com/2015/11/interpreting-fft-results-obtaining-magnitude-and-phase-information/

Example of DFT

$$F = 2kHz$$

$$f(t) = 10\sin(2\pi 10t) + 3\sin(2\pi 100t)$$



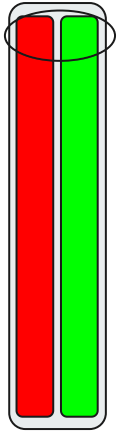
https://www.gaussianwaves.com/2015/11/interpreting-fft-results-obtaining-magnitude-and-phase-information/

Why spectrum is mirroring?

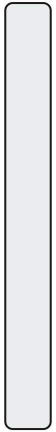
$$X_m = \sum_{n=0}^{N-1} x_n \exp\left(-j2\pi \frac{m}{N} n\right)$$

$$\begin{aligned}
X_{N-m} &= \sum_{n=0}^{N-1} x_n \exp\left(-j2\pi \frac{N-m}{N} n\right) \\
&= \sum_{n=0}^{N-1} x_n \exp\left(-j2\pi n + j2\pi \frac{m}{N} n\right) \\
&= \sum_{n=0}^{N-1} x_n \exp\left(j2\pi \frac{m}{N} n\right) \\
&= (X_m)^*
\end{aligned}$$

Discrete Fourier transform



$$\in \mathbb{C} \quad = \begin{pmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & e^{-\frac{2\pi i}{N}} & e^{-\frac{4\pi i}{N}} & e^{-\frac{6\pi i}{N}} & \dots & e^{-\frac{2\pi i}{N}(N-1)} \\ 1 & e^{-\frac{4\pi i}{N}} & e^{-\frac{8\pi i}{N}} & e^{-\frac{12\pi i}{N}} & \dots & e^{-\frac{2\pi i}{N}2(N-1)} \\ 1 & e^{-\frac{6\pi i}{N}} & e^{-\frac{12\pi i}{N}} & e^{-\frac{18\pi i}{N}} & \dots & e^{-\frac{2\pi i}{N}3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{-\frac{2\pi i}{N}(N-1)} & e^{-\frac{2\pi i}{N}2(N-1)} & e^{-\frac{2\pi i}{N}3(N-1)} & \dots & e^{-\frac{2\pi i}{N}(N-1)^2} \end{pmatrix}$$



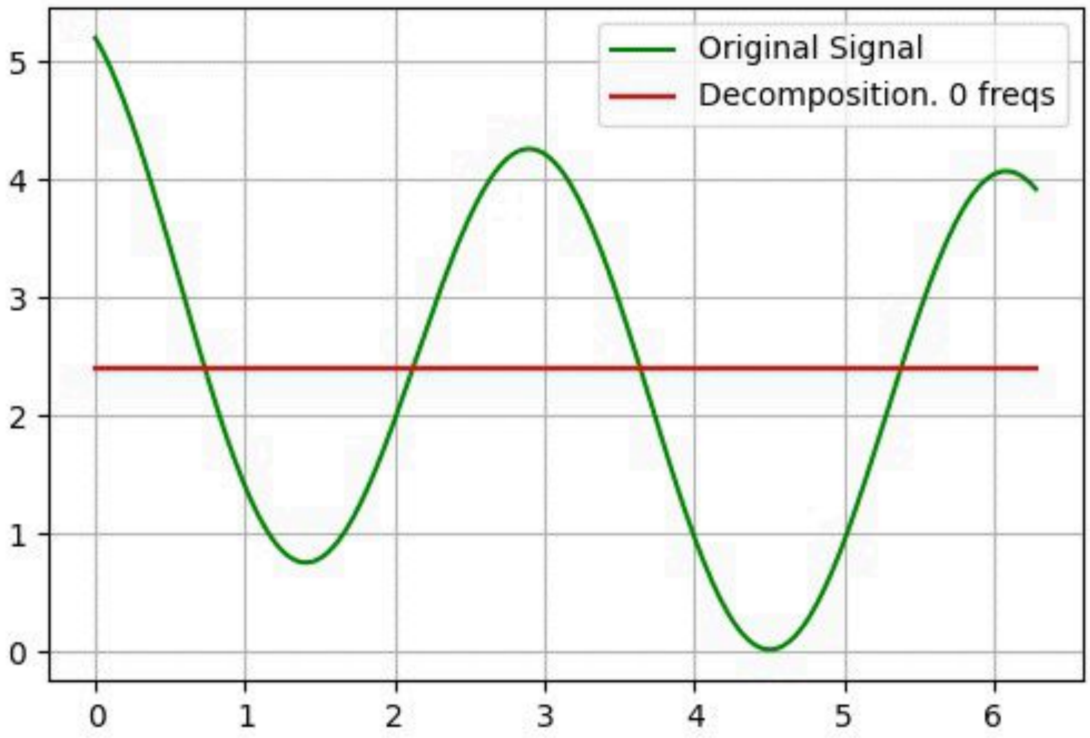
Magnitude
Phase

$$A\cos(\omega t + \phi) = B\cos(\omega t) + C\sin(\omega t)$$

$$A = \sqrt{B^2 + C^2}, \qquad \tan\phi = \frac{C}{B}$$

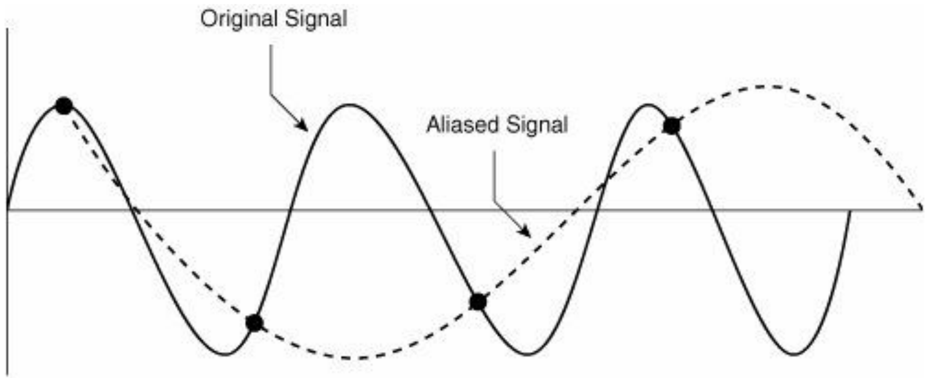
Evaluating quality of DFT

$$f(t) = 5 + 2\sin(2t + 2) - 3\cos(0.2t - 1)$$



Kotelnikov Theorem

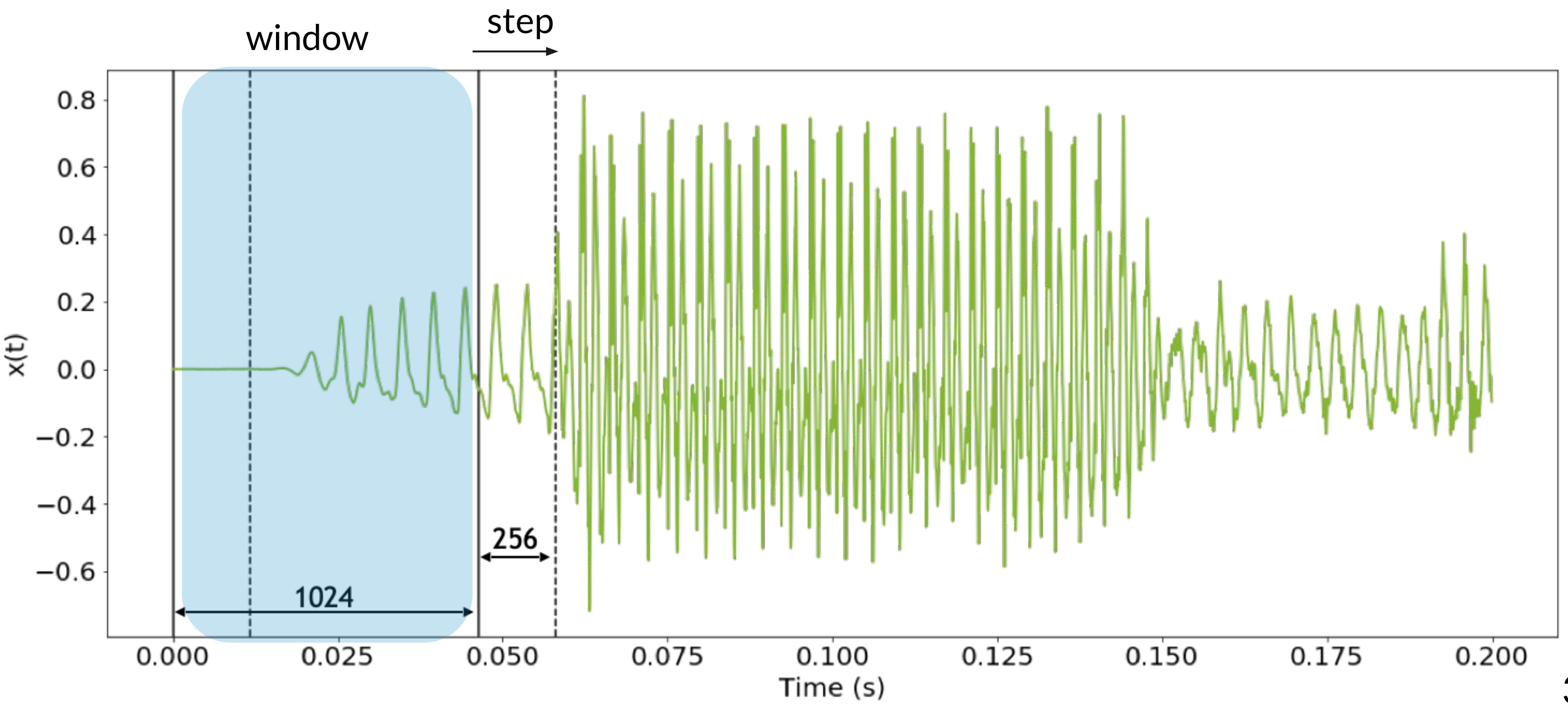
- If a function **f(t)** contain no frequencies higher than **B hertz**, it is completely determined by giving its ordinates at series of points spaced **1/2B** seconds apart
- **Example:** If signal contains frequency 100 Hz, the sampling rate for this signal needs to be 200 Hz at least
- DFT of a segment of a signal with sample rate N, will produce amplitudes for n_fft evenly spread frequencies in range [-sample_rate / 2; sample_rate /2]



Short Time Fourier Transform (STFT)

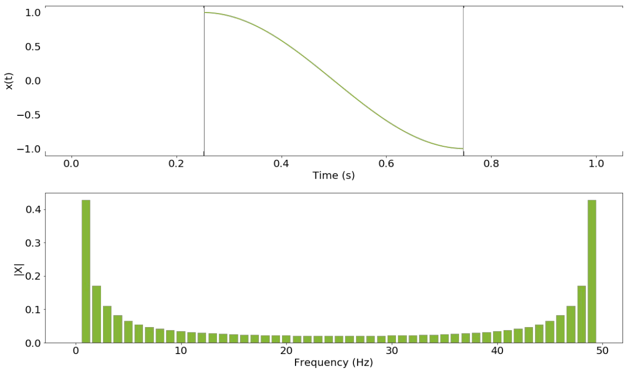
How to apply FT to a long non-periodic signal?

Short-Time Fourier Transform

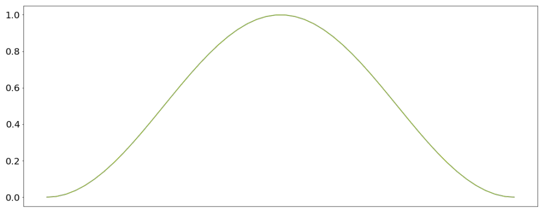


Window functions

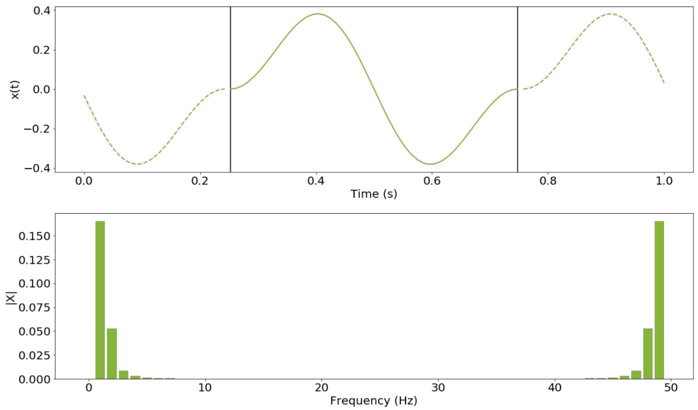
Sliced signal



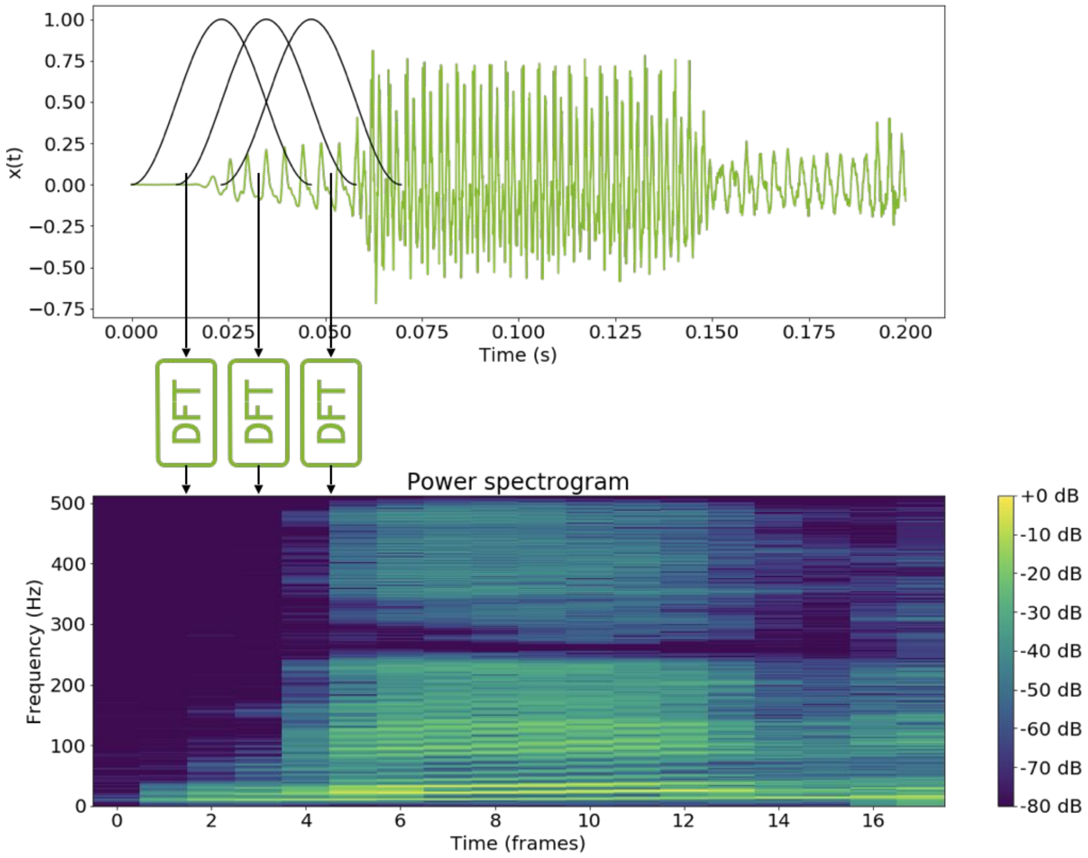
Window



Windowed signal



Short Time Fourier Transform + window function

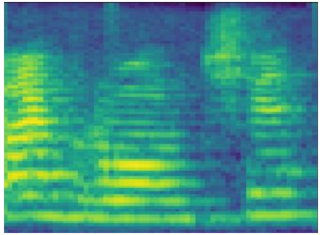
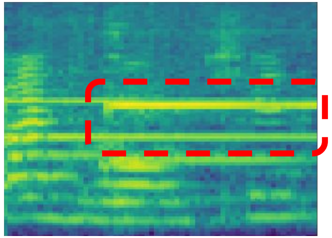


Spectrogram

Assembling everything together

- Sound representation
- Motivation for spectrograms
- Fourier Transform
- Discrete Fourier Transform
- Short Time Fourier Transform
- **Spectrogram**
- Mel scale
- MFCC

Spectrogram

	Wave	Spectrogram
Voice		
Voice + Noise		

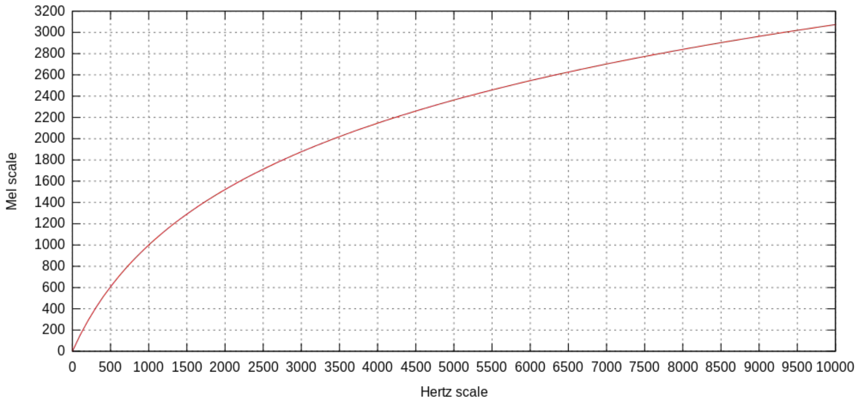
Practical use: values of the spectrogram are very small, so typically the log-spectrogram is used instead (don’t forget to add +e)

Mel Scale

Compressing the spectrogram

Mel Scale

- Humans perceive sound on a log-scale. For human ear:
  - 500 Hz << 600 Hz
  - but 5000 Hz ~= 5100 Hz



There is no single mel-scale formula.[3] The popular formula from O'Shaughnessy's book can be expressed with different logarithmic bases:

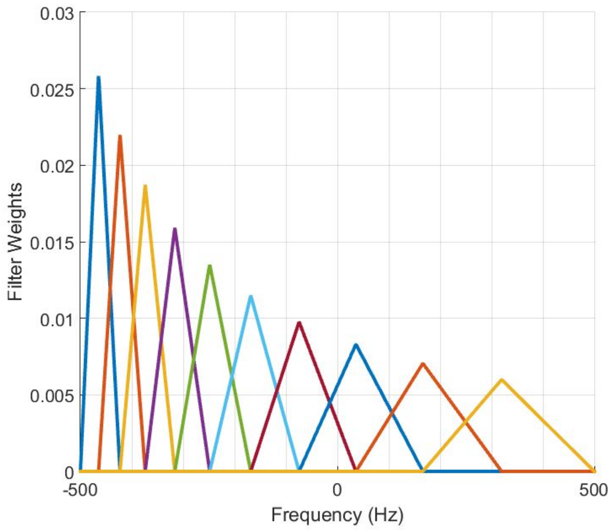
$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \ln\left(1 + \frac{f}{700}\right)$$

The corresponding inverse expressions are:

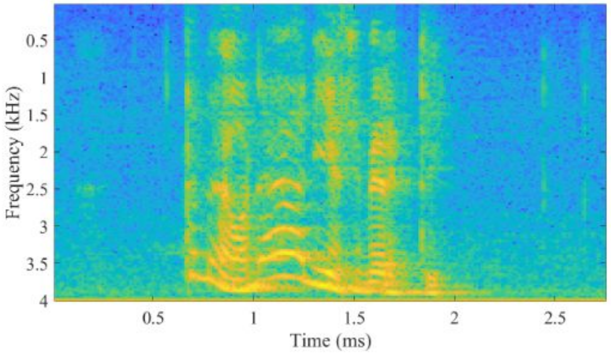
$$f = 700\left(10^{\frac{m}{2595}} - 1\right) = 700\left(e^{\frac{m}{1127}} - 1\right)$$

https://en.wikipedia.org/wiki/Mel_scale

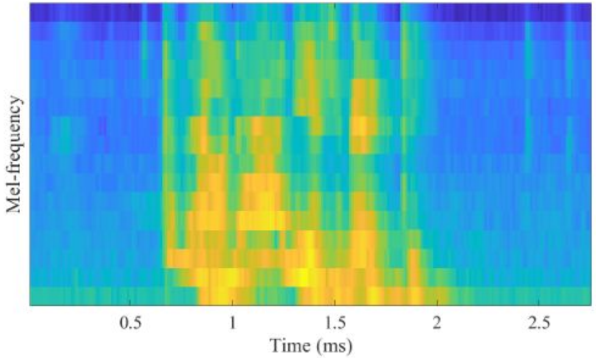
Mel Spectrogram



Spectrogram of a segment of speech



Spectrogram after multiplication with mel-weighted filterbank



https://wiki.aalto.fi/display/ITSP/Cepstrum+and+MFCC

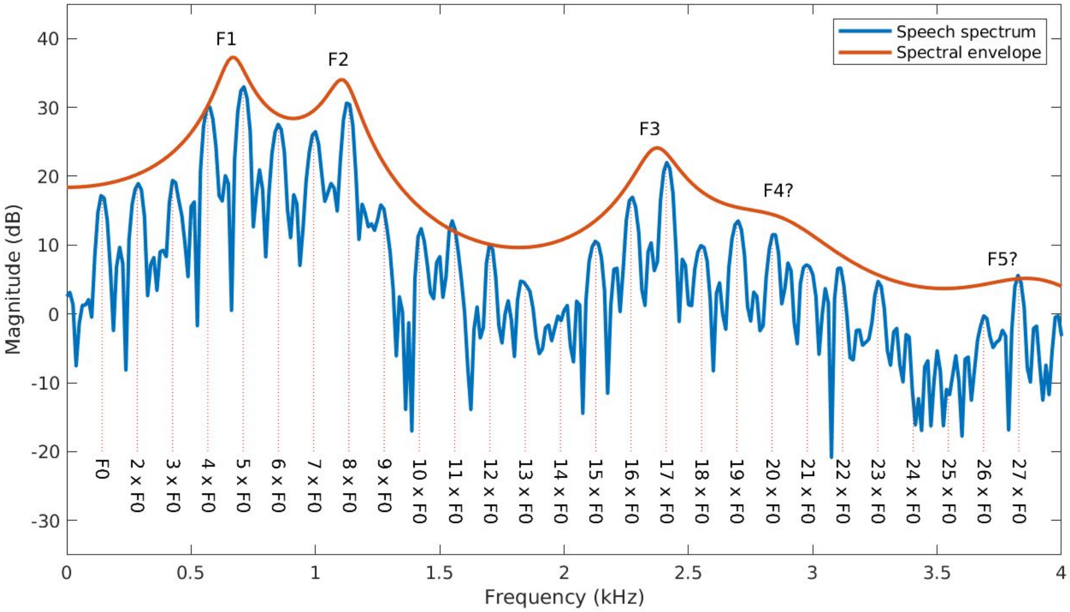
MFCC

Decorrelating the spectrogram

- Sound representation
- Motivation for spectrograms
- Fourier Transform
- Discrete Fourier Transform
- Short Time Fourier Transform
- Spectrogram
- Mel scale
- **MFCC**

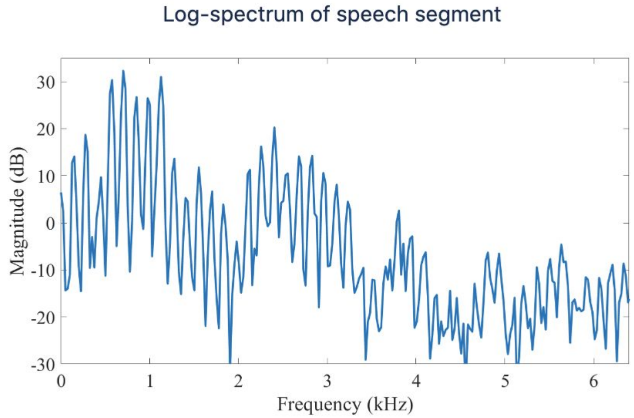
Fundamental Frequency

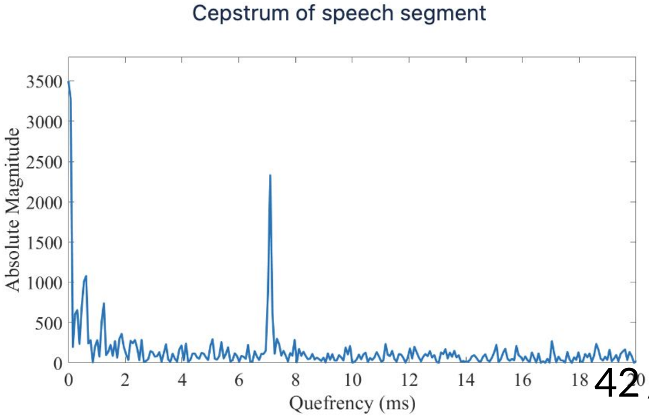
- **Fundamental frequency** refers to the approximate frequency of the (quasi-)periodic structure of voiced speech signals
- Peaks on envelope curve are **formants**
- **Pitch** is perceptual value, F0 is physical
- F0 lie roughly in the **range 80 to 450 *Hz***, where males have lower voices than females and children



Cepstrum

- Fourier spectrum of voice has **periodic** structure
- Apply **DCT** (Discrete Cosine Transform) to spectrum and obtain **Cepstrum**
- **Peak** in Cepstrum should be located at $\dfrac{1}{F_0}$



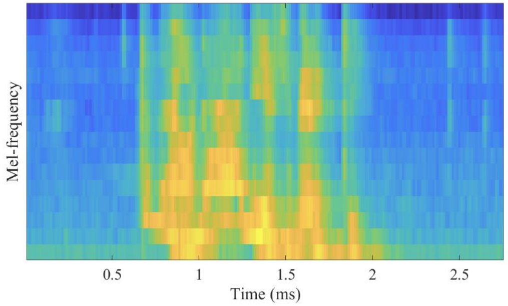


https://wiki.aalto.fi/display/ITSP/Cepstrum+and+MFCC

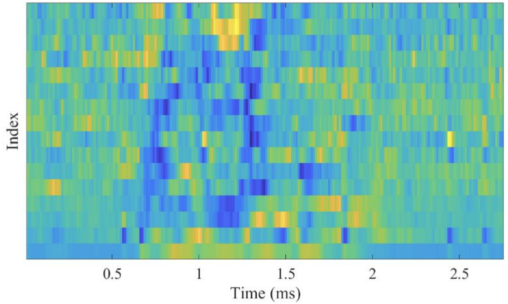
Mel-Frequency Cepstral Coefficients (MFCCs)

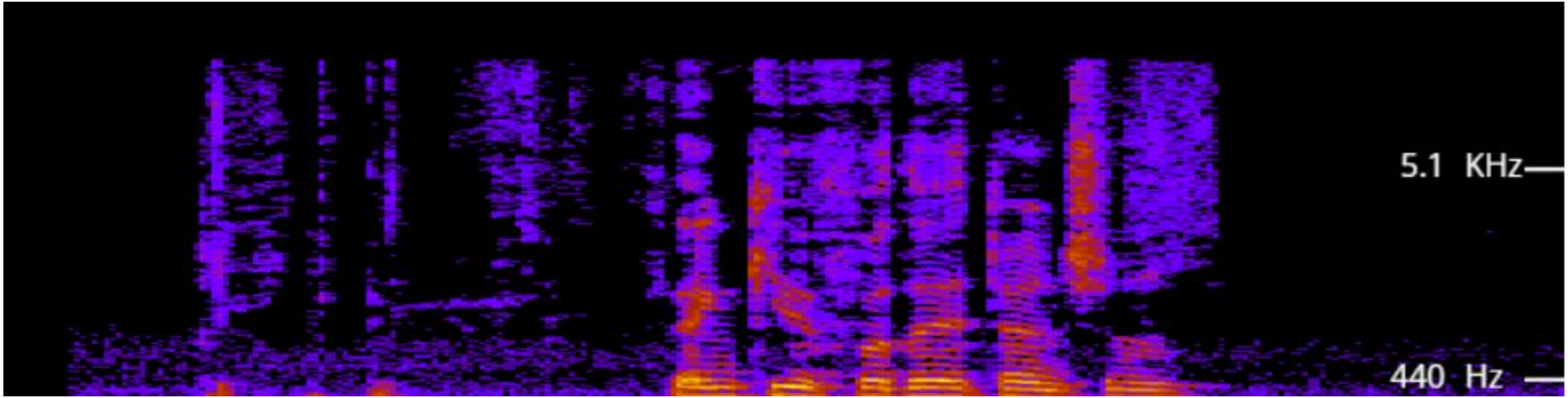
- Algorithm of acquiring MFCC:
    - Apply STFT to the signal
    - Apply mel filters
    - Take the log value
    - Apply DCT

Spectrogram after multiplication with mel-weighted filterbank



Corresponding MFCCs




5.1 KHz
440 Hz