Introduction to DSP

A short history of Speech Recognition

50's

In **1952**, *Bell Laboratories* designed the "**Audrey**" system which could recognize a single voice speaking **digits** aloud

In **1962,** *IBM* introduced "**Shoebox**" which understood and responded to **16 words** in English.

60's

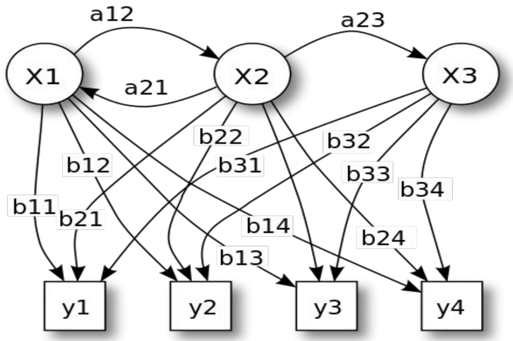
A short history of Speech Recognition

70's

DARPA's system was capable of understanding over **1,000** words. **Siri** was a spin-out of DARPA development :)

80's

The '80s saw speech recognition vocabulary go from a few hundred words to **several thousand words** thanks to **HMM**



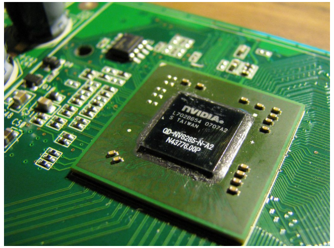
A short history of Speech Recognition

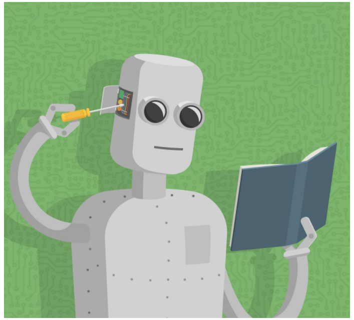
90's

Speech recognition was propelled forward in the 90s in large part because of **faster processors**

00-10's

And then came the era of big data, machine learning and GPUs





A short history of Speech Synthesis

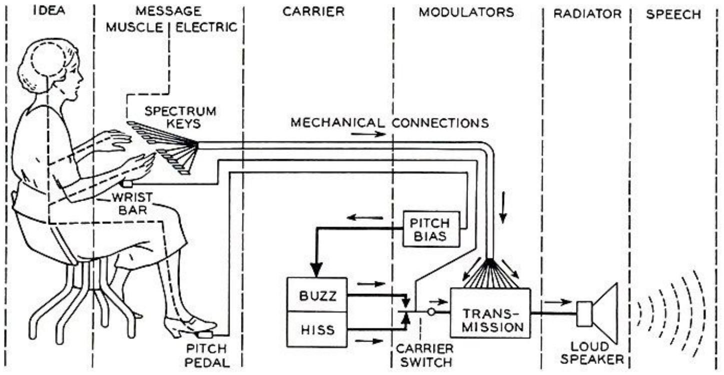
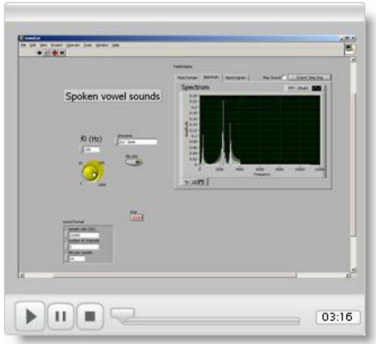


Fig. 8—Schematic circuit of the voder.

30's

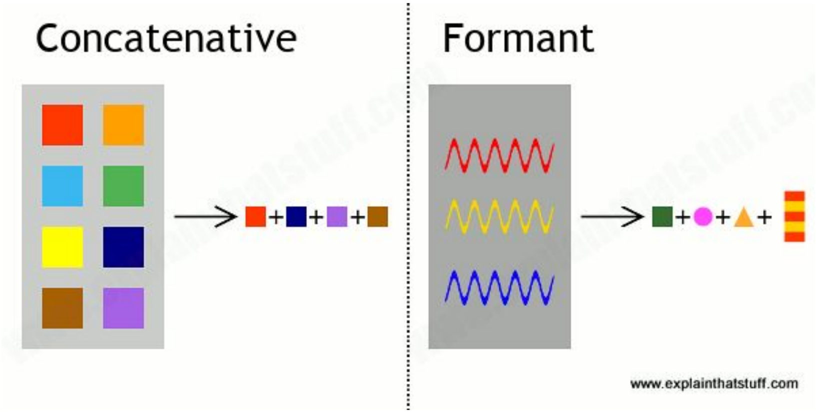
until 80's

In **1939**, *The Bell Laboratory*'s **Voder** was the first attempt to electronically synthesize human speech by breaking it down into its **acoustic components**



Formant-based on rules. You may listen examples in Atari&Sega games :)

A short history of Speech Synthesis

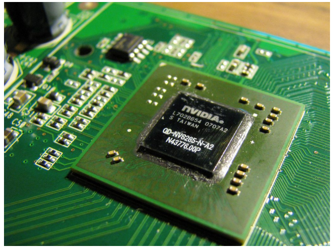


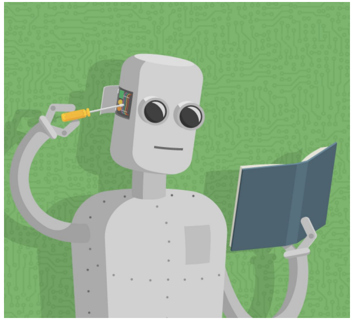
90-00's

Concatenative synthesis is a technique for synthesising sounds by concatenating short samples of recorded sound (called *units*).

And then came the era of big data, machine learning and GPUs

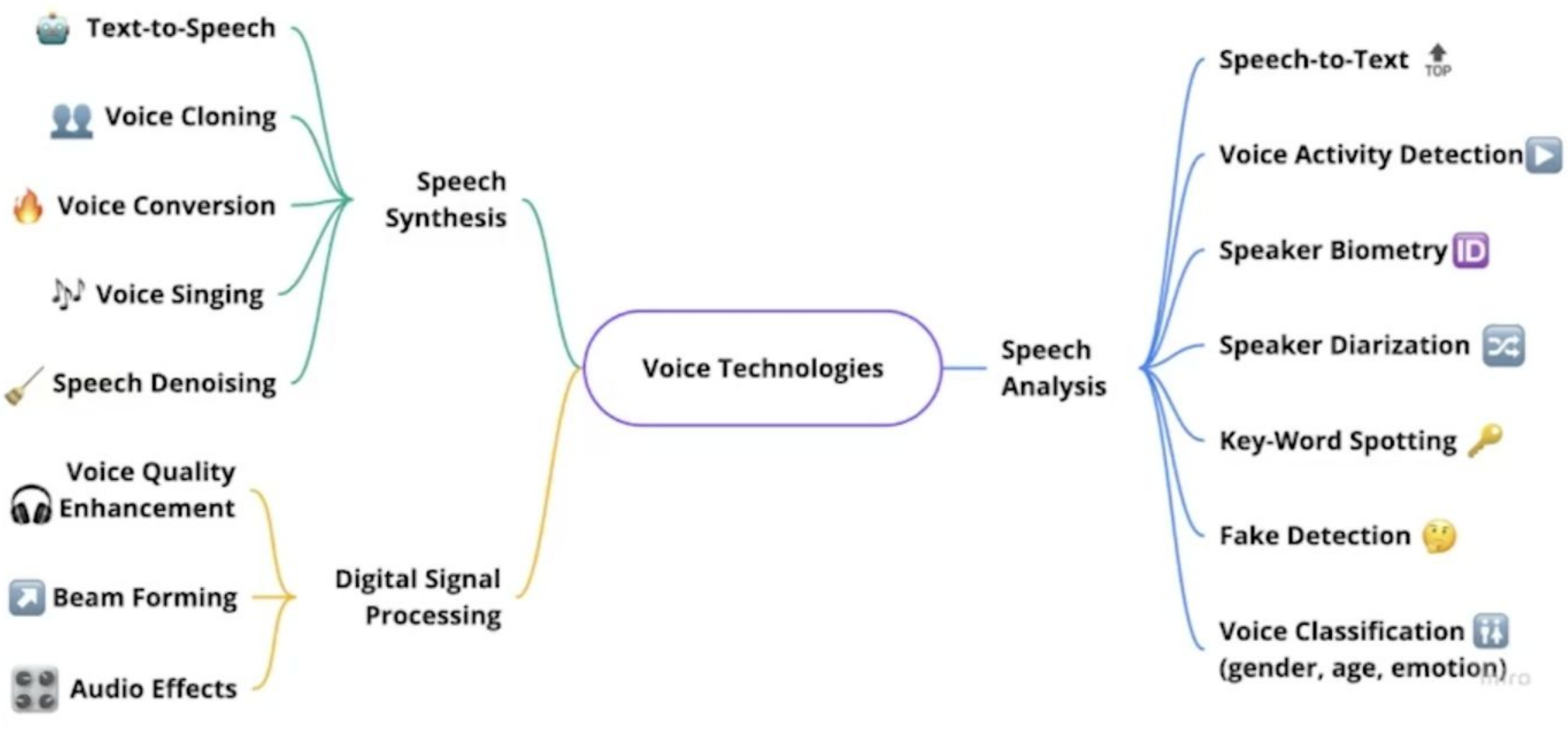
10's





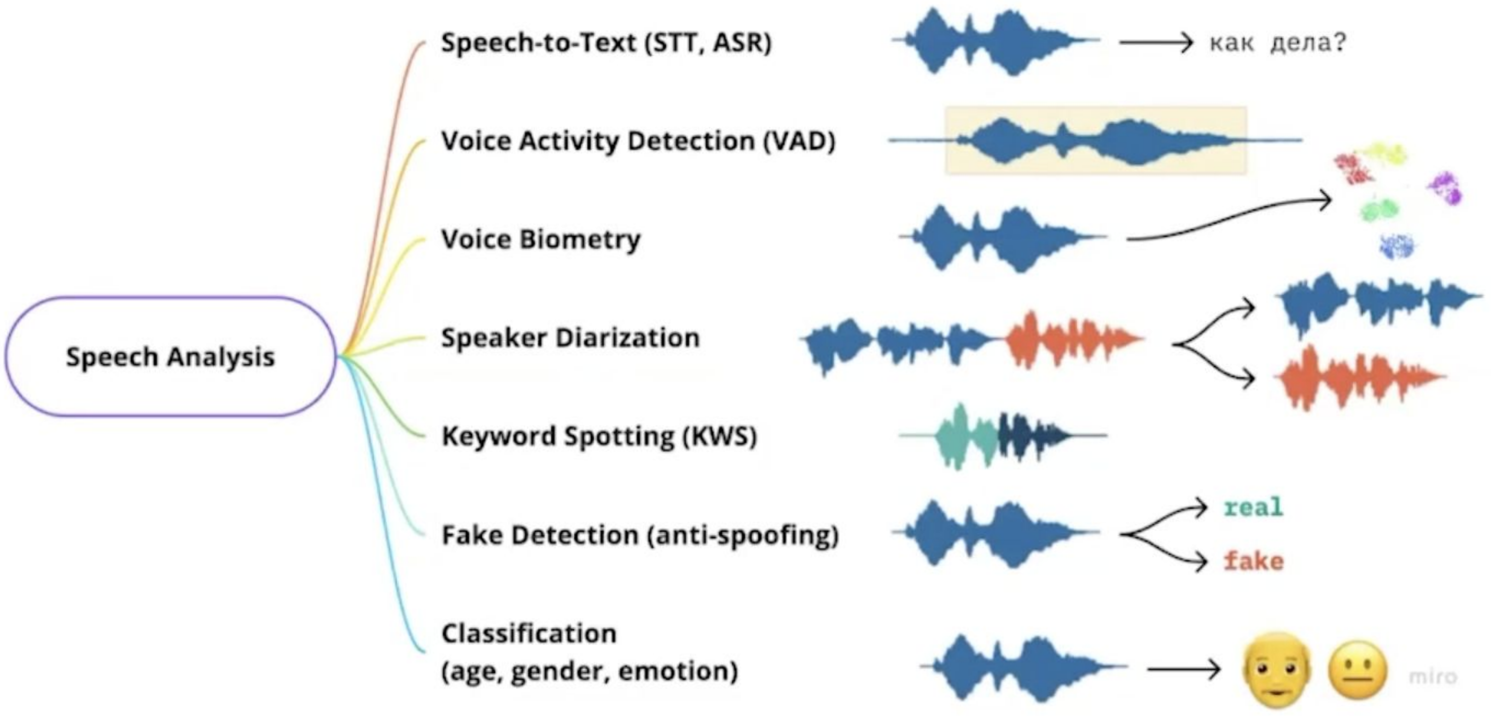
Voice Technologies Applications

Voice Technologies Mind Map



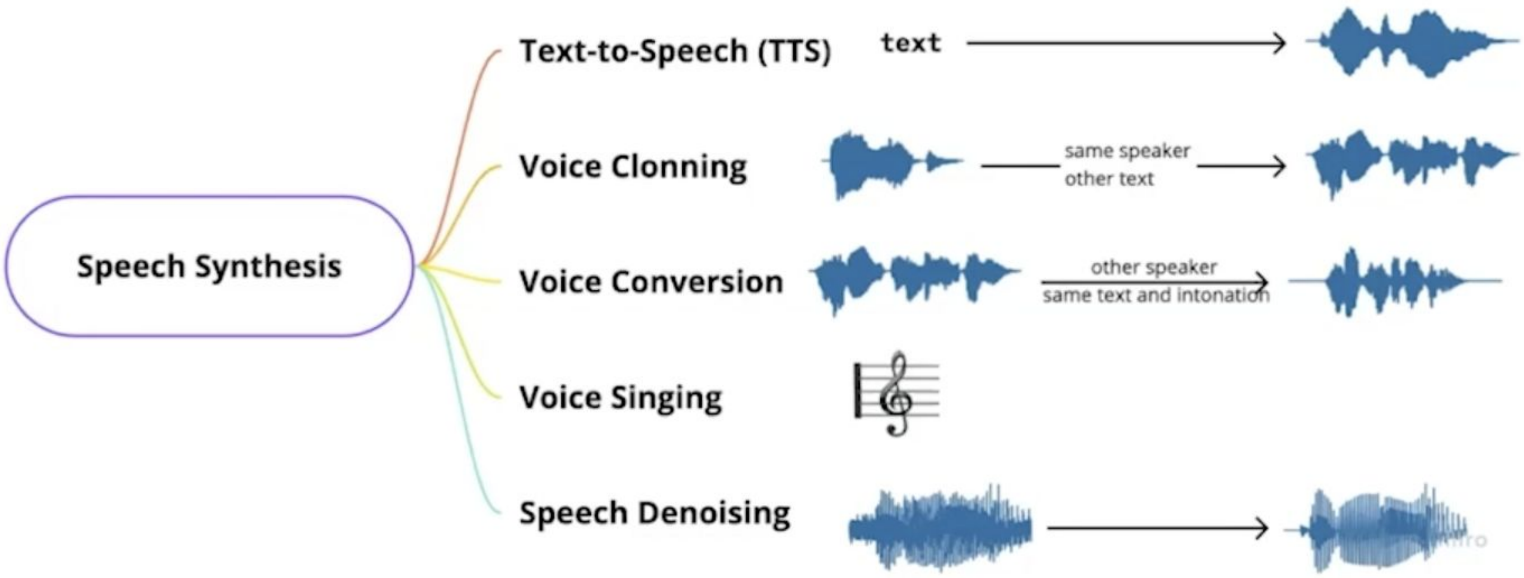
https://github.com/yandexdataschool/speech_course

Voice Technologies Mind Map



https://github.com/yandexdataschool/speech_course

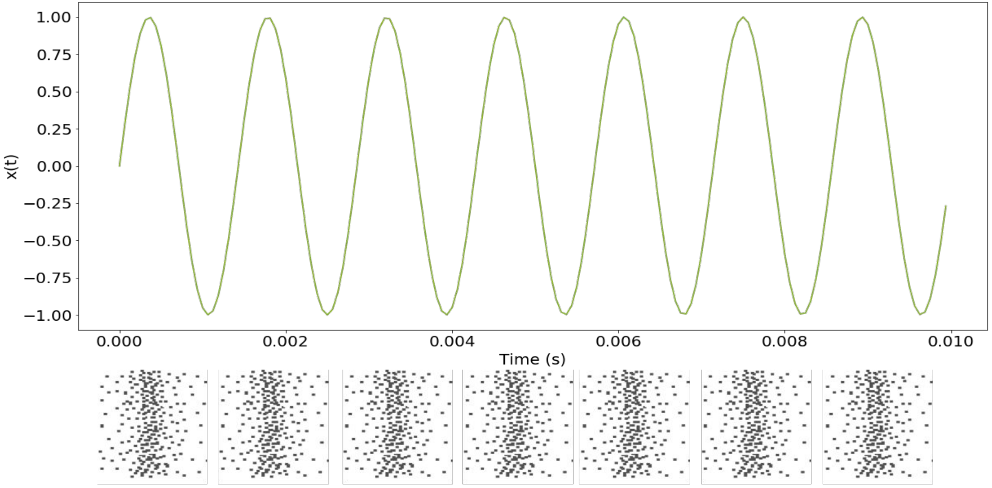
Voice Technologies Mind Map

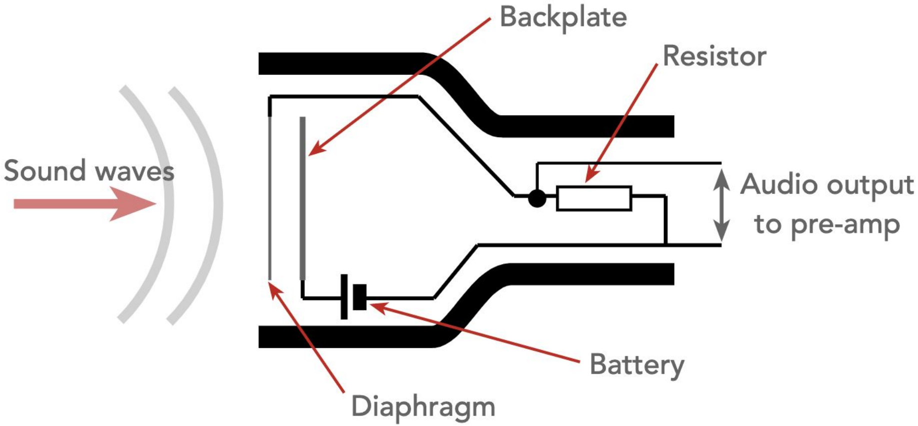


https://github.com/yandexdataschool/speech_course

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

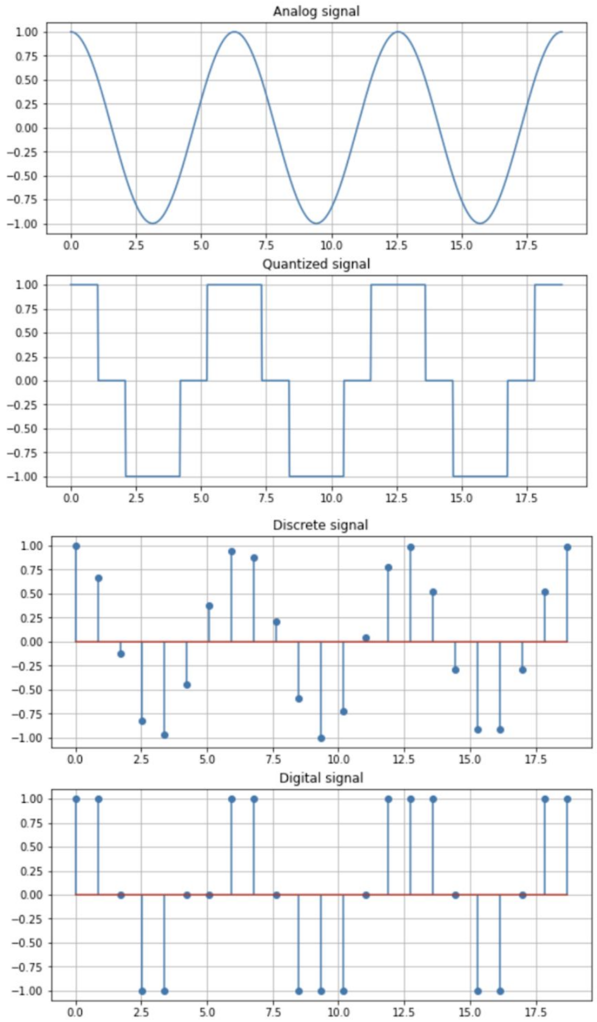




https://pudding.cool/2018/02/waveforms/
https://blog.accusonus.com/science-of-sound/how-microphones-work

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

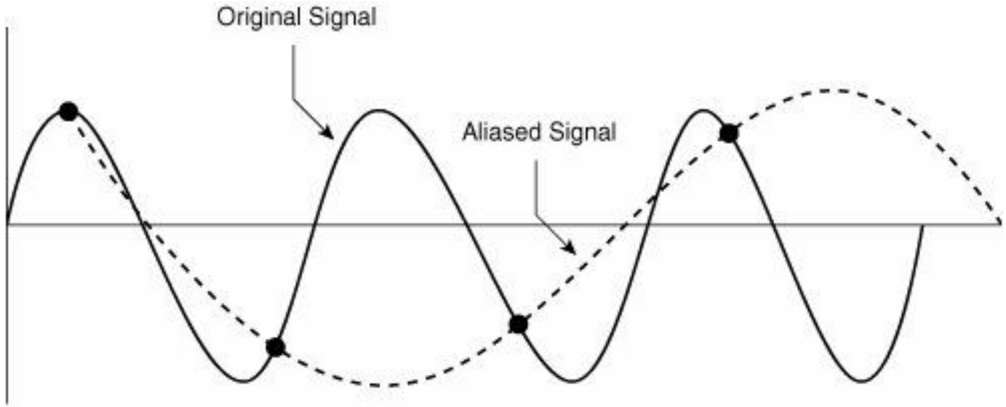


https://github.com/markovka17/dla/blob/master/week01/dsp.ipynb
https://web.sonoma.edu/esee/courses/ee442/archives/sp2019/lectures/lecture09_pcm.pdf

Kotelnikov Theorem

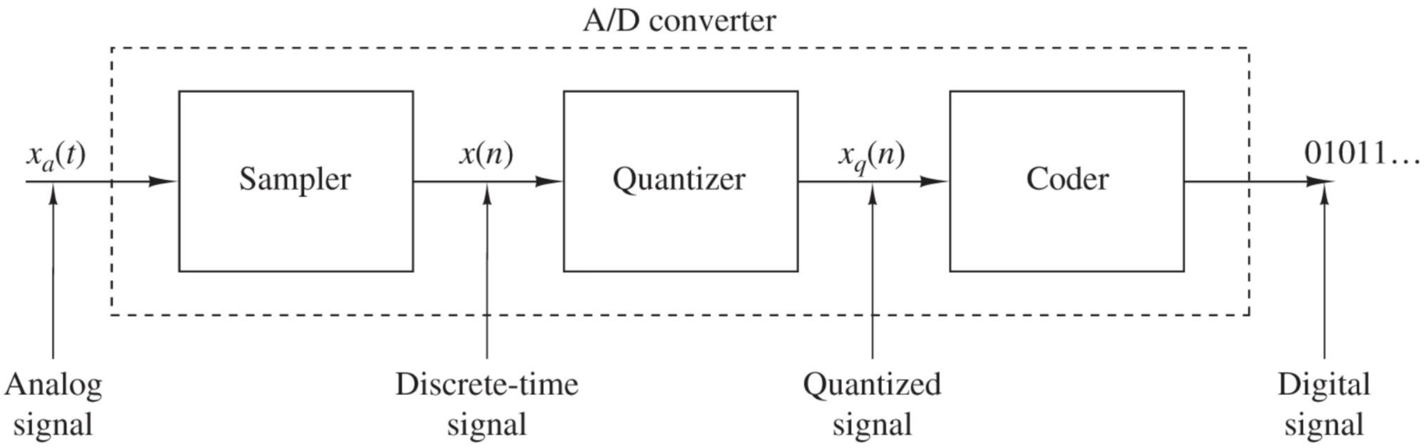
- If a function **f(t)** contain no frequencies higher than **B hertz**, it is completely determined by giving its ordinates at series of points spaced **1/2B** seconds apart

- **Example:** If signal contains frequency 100 Hz, the sampling rate for this signal needs to be 200 Hz at least

- 

Analog-to-Digital Conversion

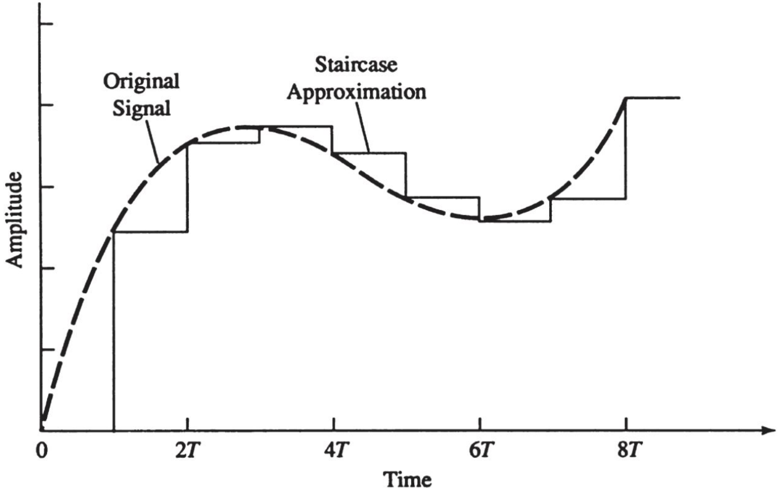
- Converting analog signals to a sequence of numbers having finite precision
- Corresponding devices are called A/D converters (ADCs)



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.

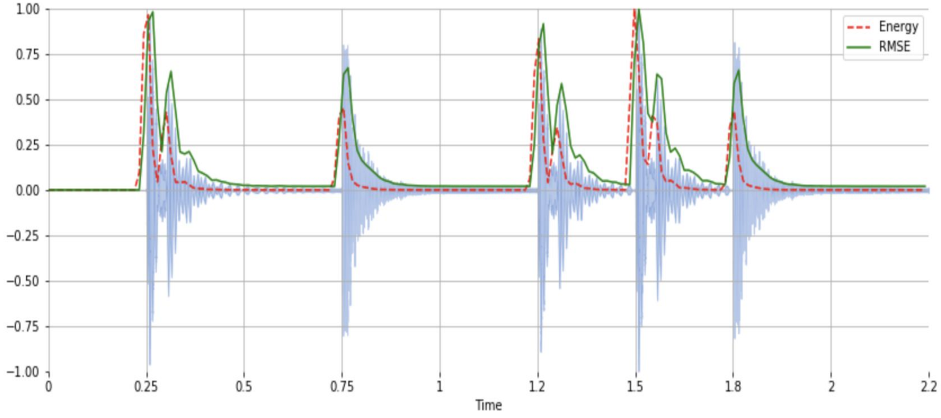
192000 Hz

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?

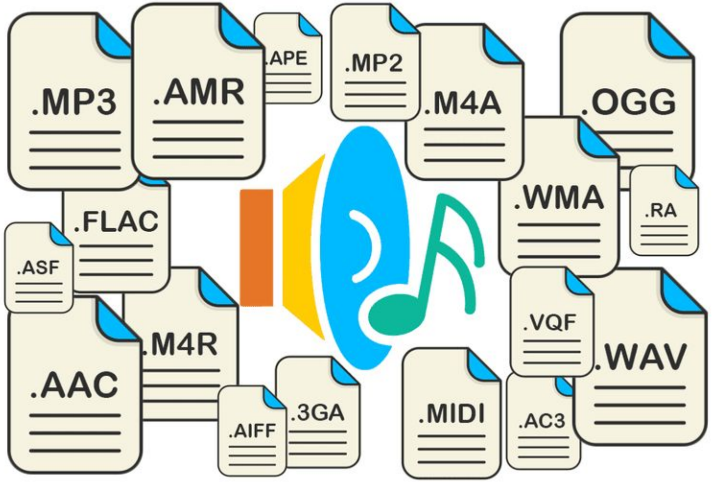
- Assume **f(n)** is our signal where **n** is time
- Power of signal is $f^2(n)$
- Energy of signal is $\sum f^2(n)$
- In practice estimated by some **window**
- Energy in **decibels:** $10 \log_{10} E$
- $\mathrm{SNR}_{dB} = 10 \log_{10} \frac{E_{\text{signal}}}{E_{\text{noise}}}$



https://musicinformationretrieval.com/energy.html

What about audio formats?

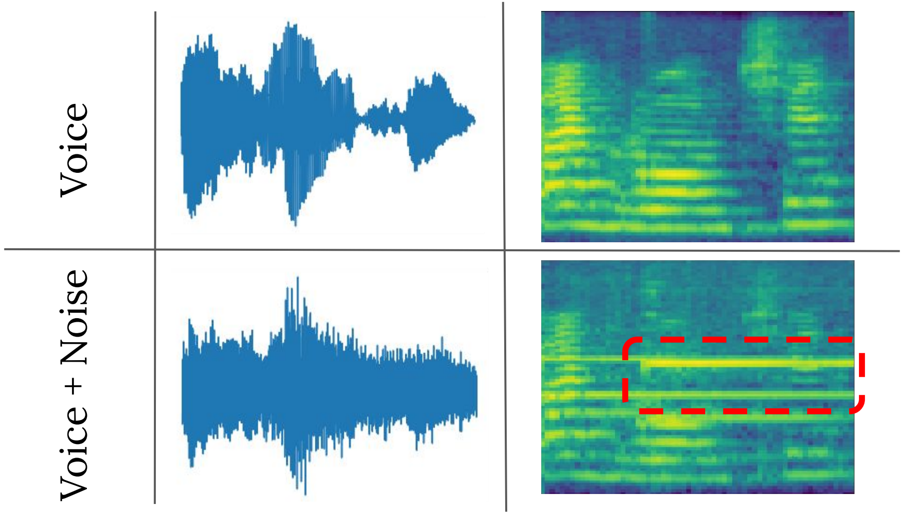
- Non-compressed formats: **WAV, AIFF, etc.**
- Lossless compression(2:1) : **FLAC, ALAC, etc.**
- Lossy compression(10:1) : **MP3, Opus, etc**
- **Bit rate** measure a degree of compression. Number of bit that are conveyed or processed per **unit of time.**

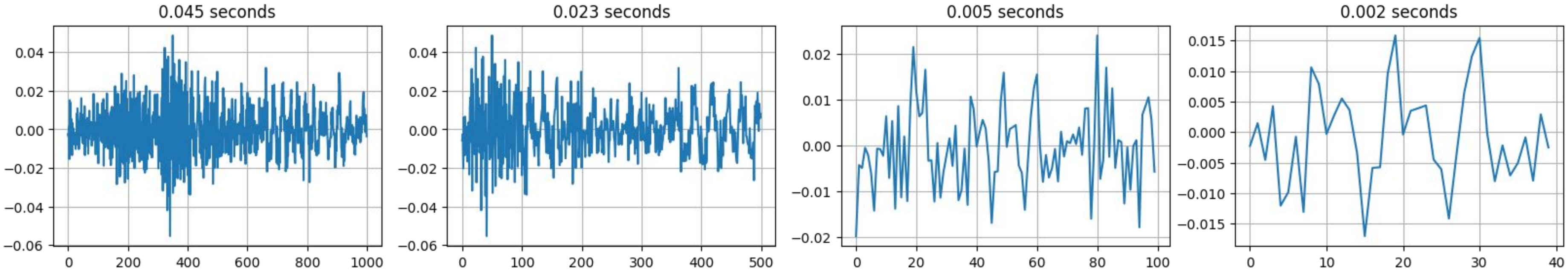


https://en.wikipedia.org/wiki/Audio_file_format

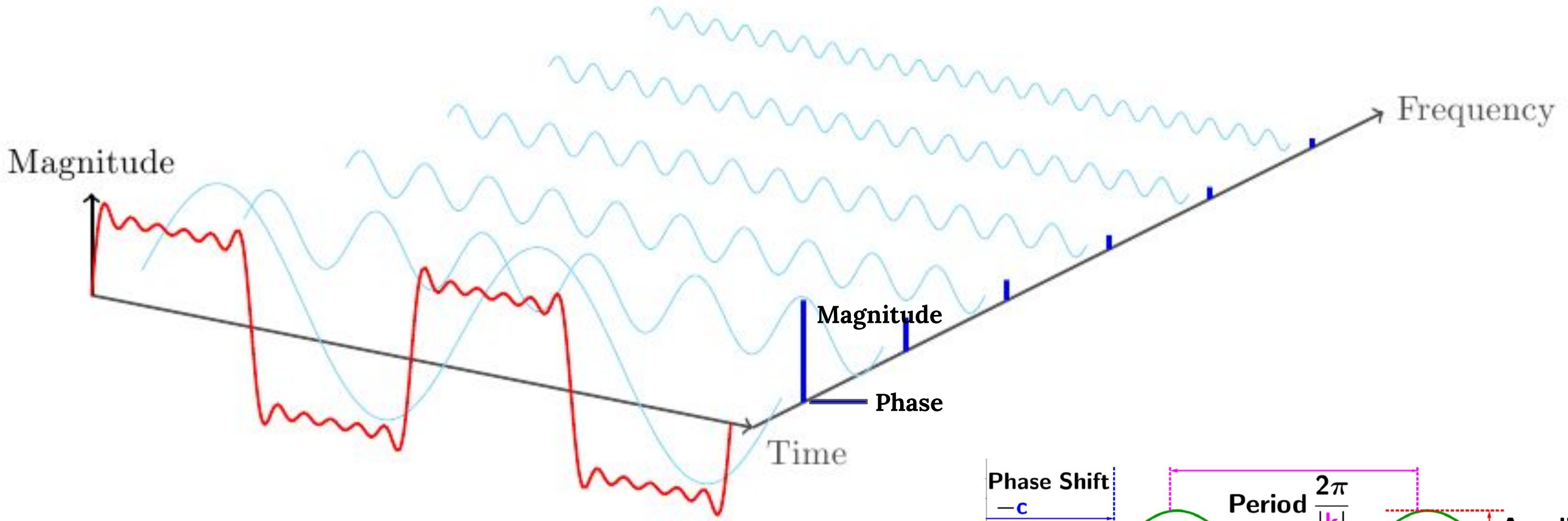
Why is it bad to work with sound in this format?

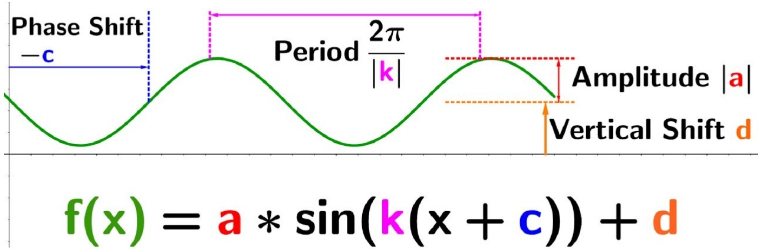
- No "invariant" regarding noise and transformations
- One letter/sound consists of 2000-4000 amplitudes, so they are expensive to process and store
- Periodical nature of audio signals





Decompose into periodic basis





$$
\begin{aligned}
A\cos(\omega t+\phi) &= A[\cos(\omega t)\cos(\phi)-\sin(\omega t)\sin(\phi)] \\
&= B\cos(\omega t)+C\sin(\omega t)
\end{aligned}
$$

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

Complex functions basis

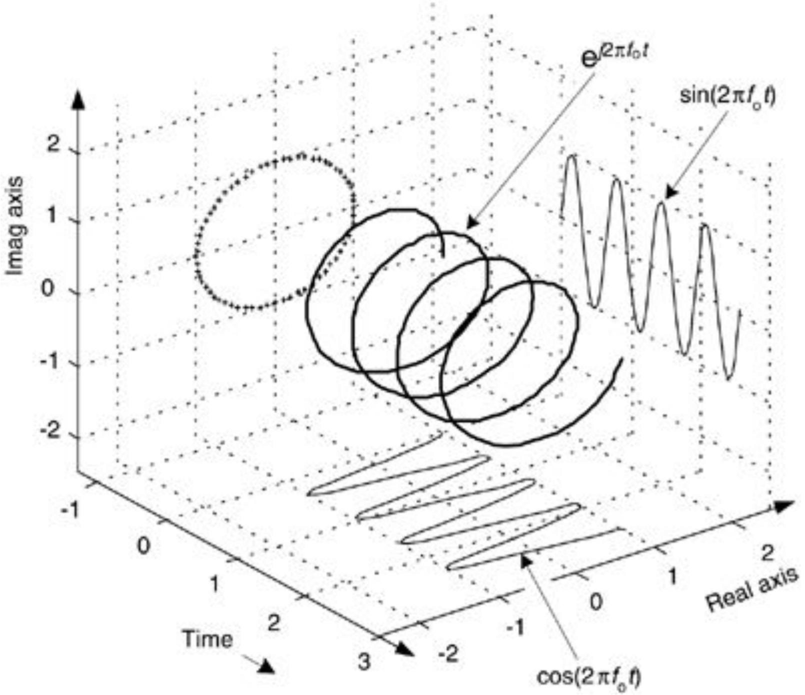
- Euler's formula

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

- Complex exponential basis

$$e^{-jwx}, w \in \mathbb{C}$$

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



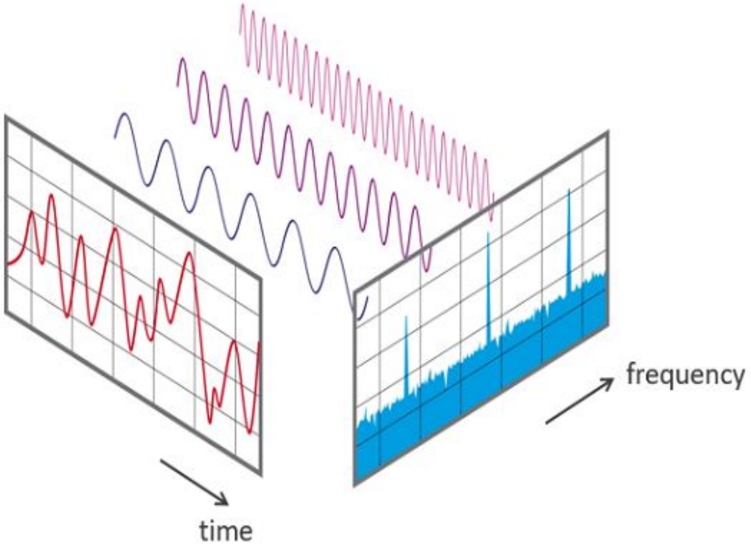
Fourier Transform

- 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 the **time domain** to the **frequency domain**

- $$F(y) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x y} dx$$

$$\text{time} \rightarrow \text{frequency}$$



https://youtu.be/spUNpyF58BY

Discrete Fourier transform

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

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

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

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

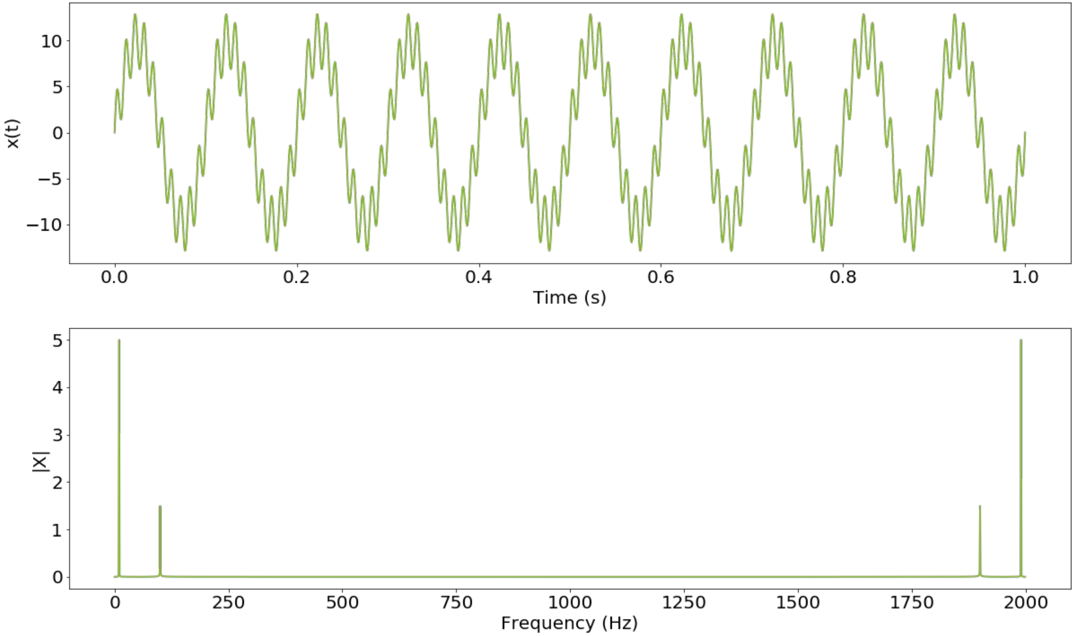
$$A\cos(\omega t + \phi) = B\cos(\omega t) + C\sin(\omega t)$$

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

Example of DFT

$$F = 2kHz$$

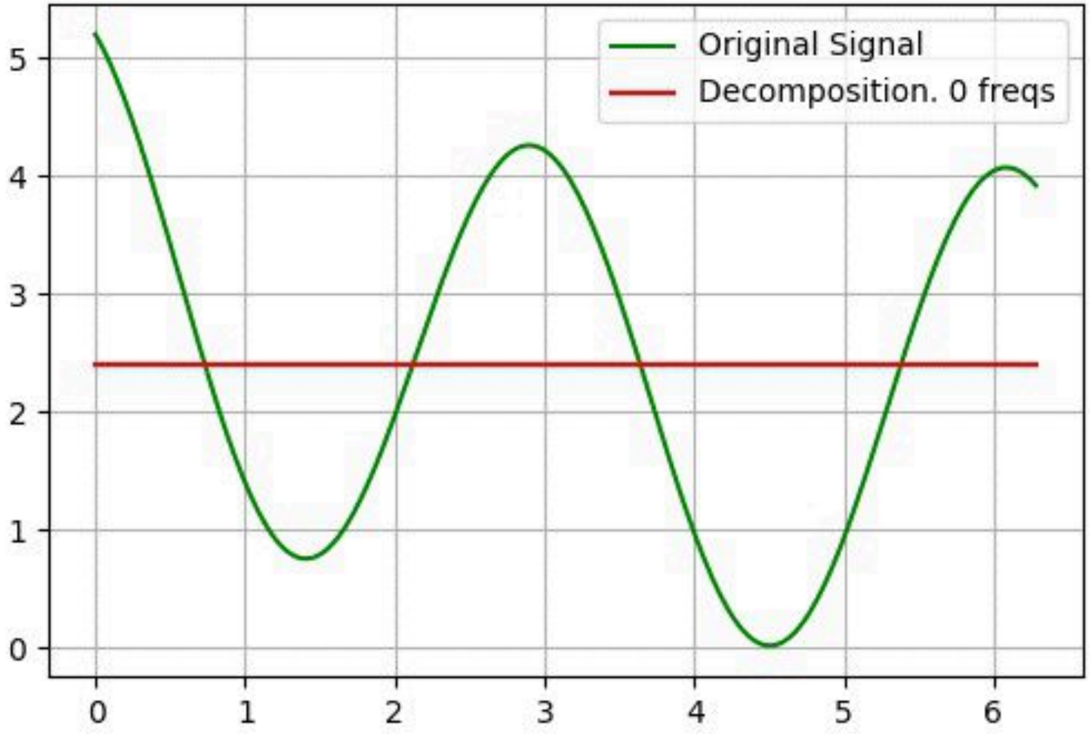
$$f(t) = 10\sin(2\pi 10t) + 3\sin(2\pi 100t)$$



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

Example of DFT

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



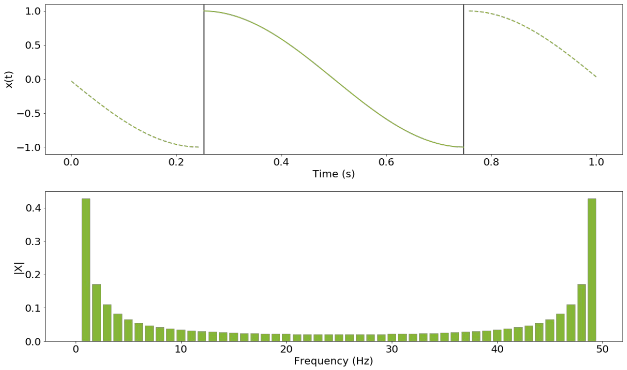
Why spectrum is mirroring?

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

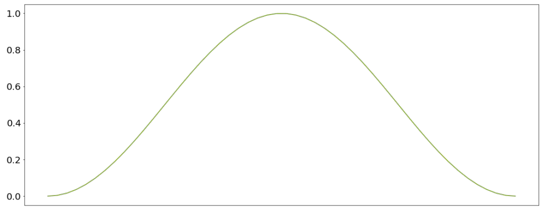
Short-time Fourier transform

FFT + Windowing

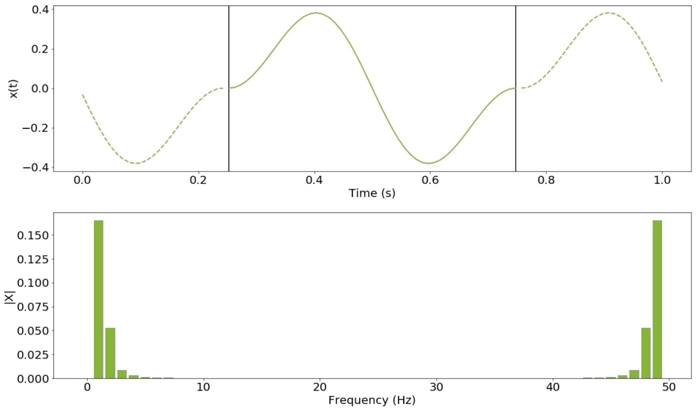
Sliced signal



Window

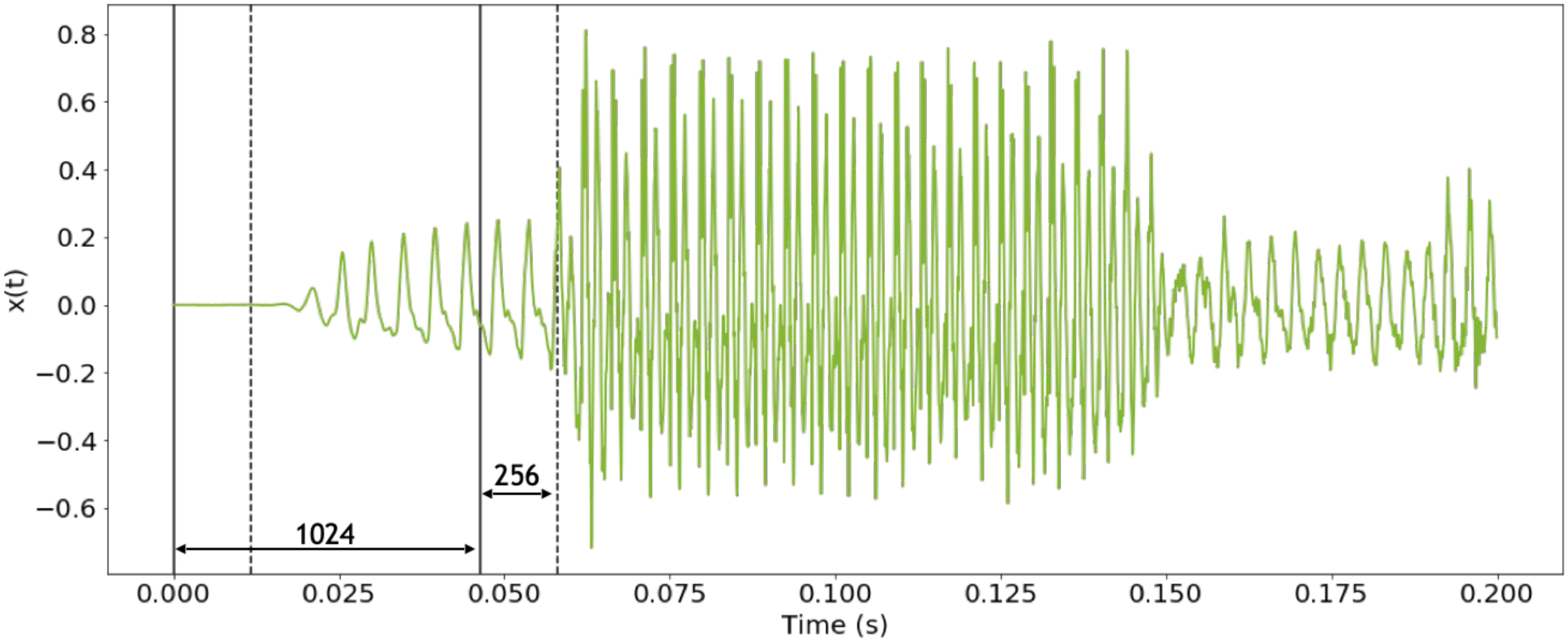


Windowed signal

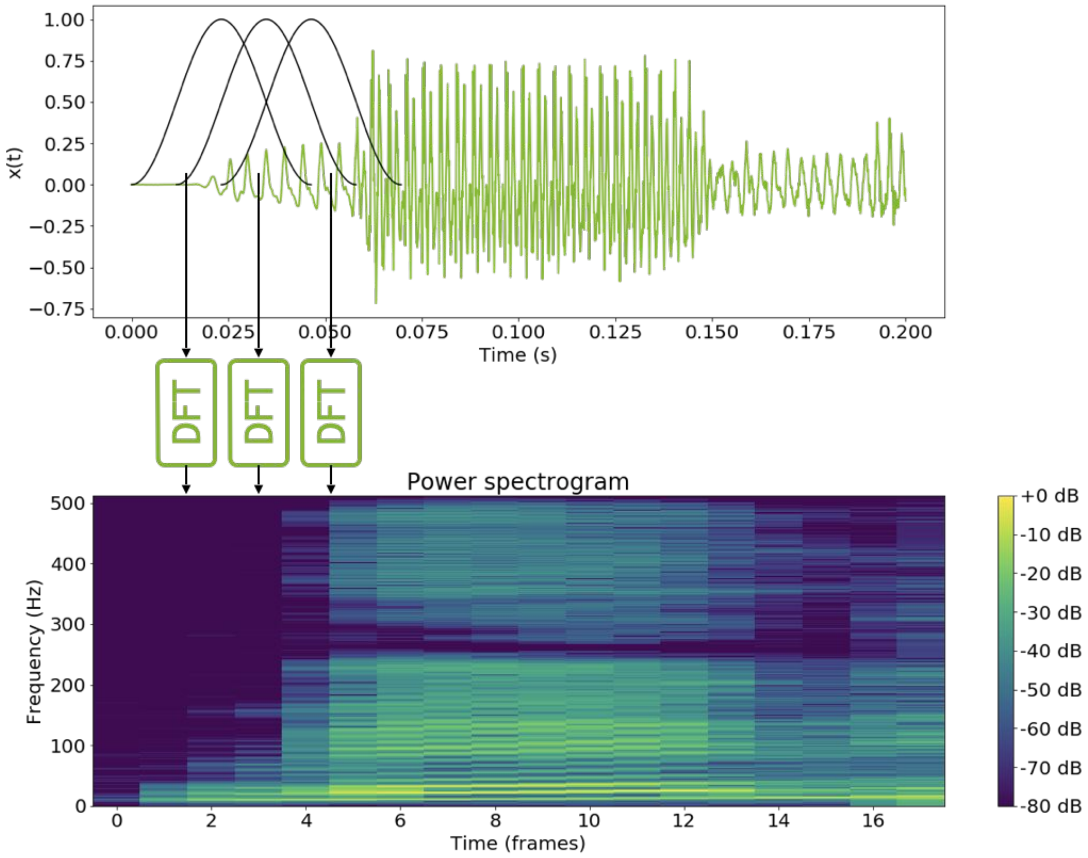


Short-time Fourier transform

FFT + Windowing



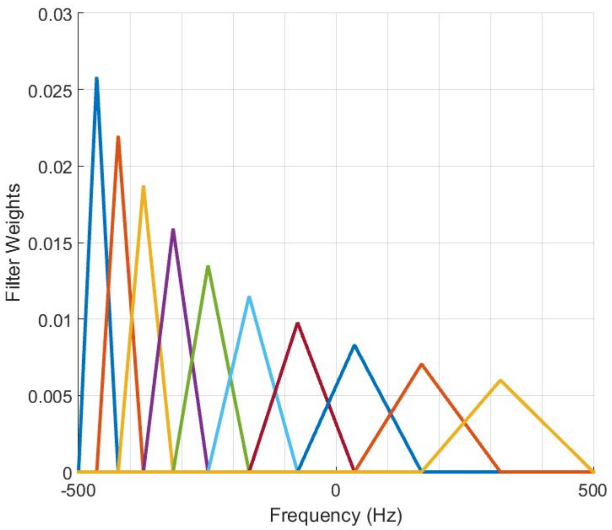
Spectrograms

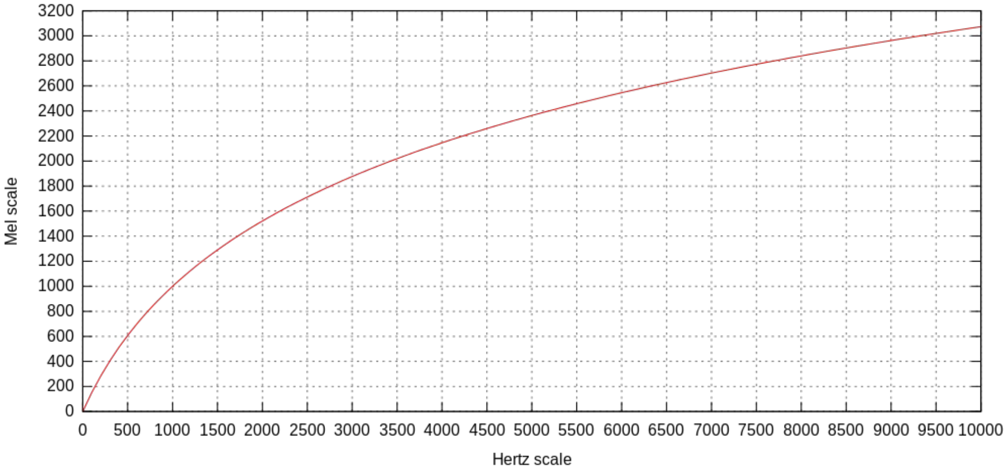


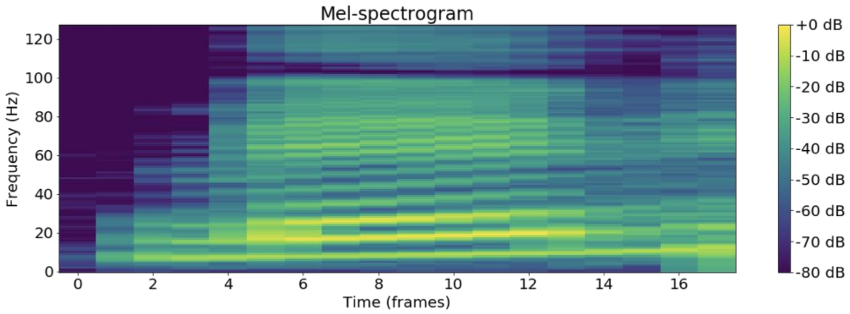
Mel Scale

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

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





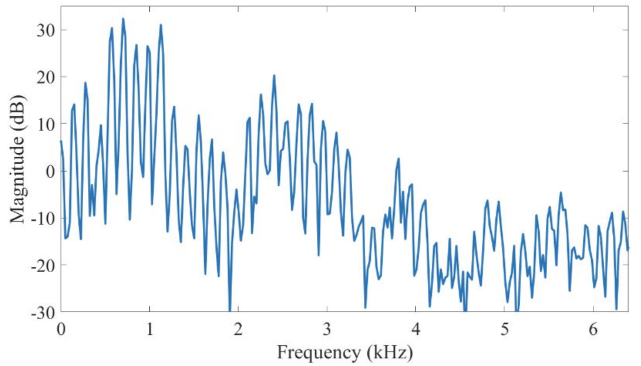


Cepstrum

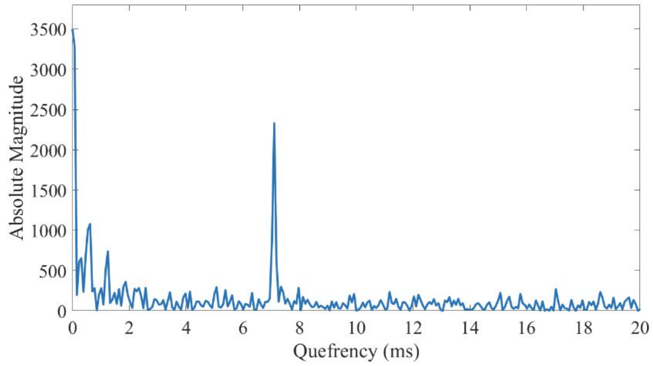
- Fourier spectrum of voice has **periodic** structure
- Apply **DCT** (Discrete Cosine Transform) to spectrum and obtain **Cepstrum**
- **Peak** in Cepstrum should be located at $\frac{1}{F_0}$

$$\text{power cepstrum of signal} = \left|\mathcal{F}^{-1}\left\{\log\left(|\mathcal{F}\{x(t)\}|^2\right)\right\}\right|^2$$

Log-spectrum of speech segment



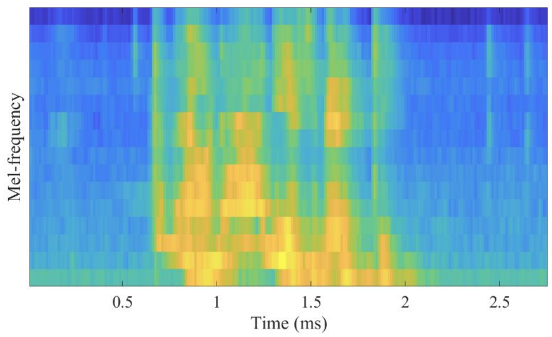
Cepstrum of speech segment



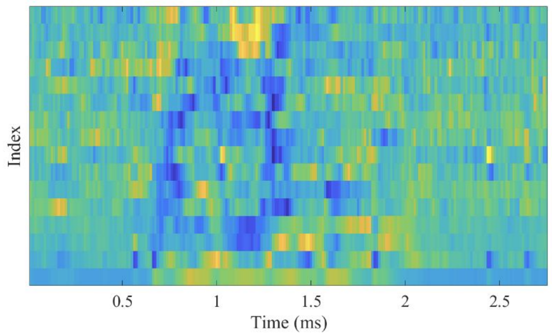
https://wiki.aalto.fi/display/ITSP/Cepstrum+and+MFCC

Mel-Frequency Cepstral Coefficients (MFCCs)

Spectrogram after multiplication with mel-weighted filterbank



Corresponding MFCCs



Pros:

- Easy to calculate
- Extracts ‘correct’ frequencies

Cons:

- Not robust to noise
- No theoretical motivation
- Don’t work for synthesis