Deep Learning for Audio

Lecture 1

Pavel Severilov

AI Masters

2025

Outline

1. Voice Technologies: tasks
2. Speech Recognition: history, basics
3. Speech Synthesis: history, basics
4. Sound: characteristics
5. Fourier Transform
6. Spectrograms

Organisation

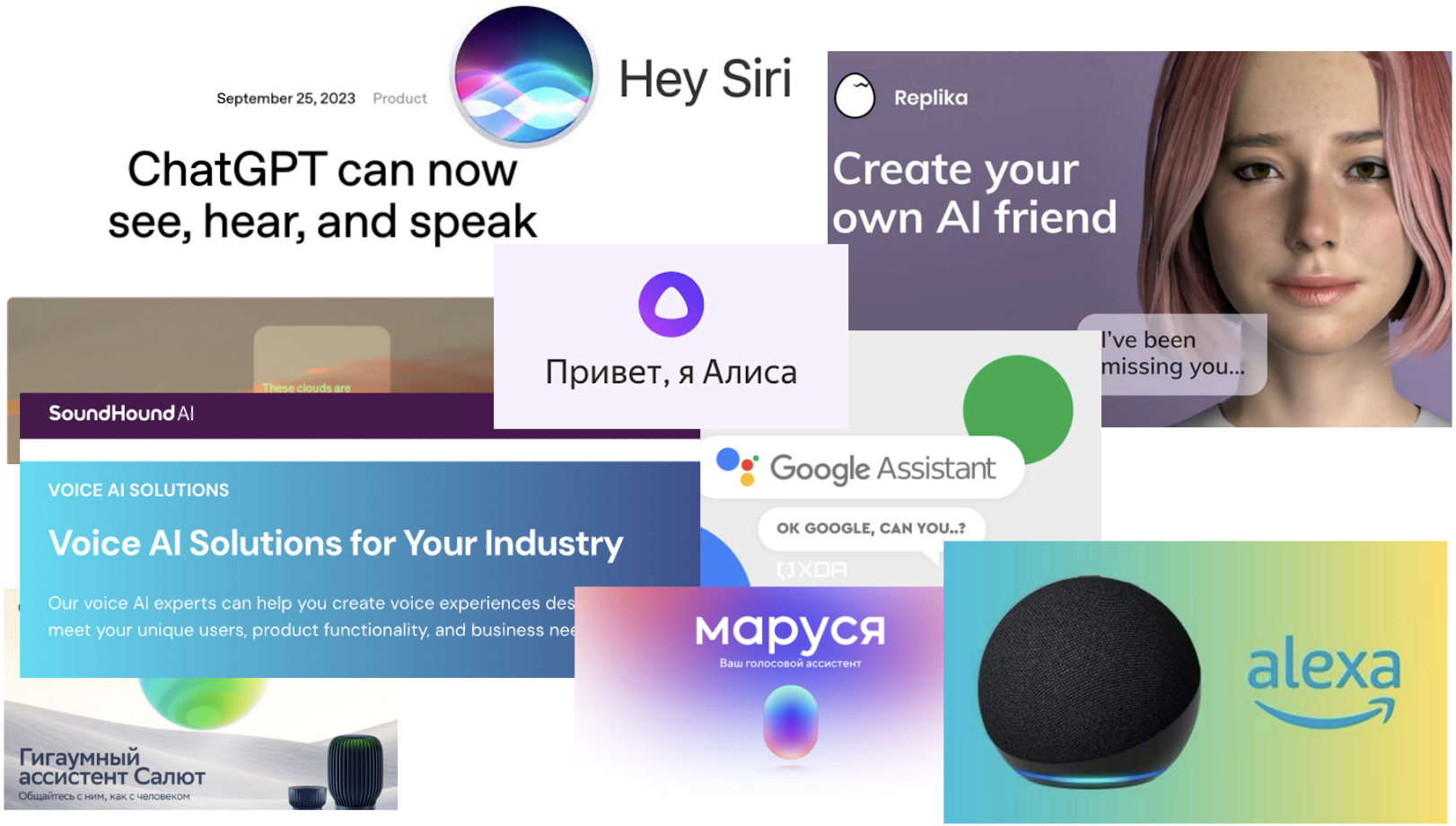
1. 10 lectures + 10 seminars
2. 4 homeworks (2 points for every)
3. Optional project (2 points)
4. hw after deadline: points divided by 2
5. Final test (2 point)
6. Grades: your points (4*2 + 2 [+ 2]= 10 [12])
7. Discussion: telegram chat

What you need to know to start

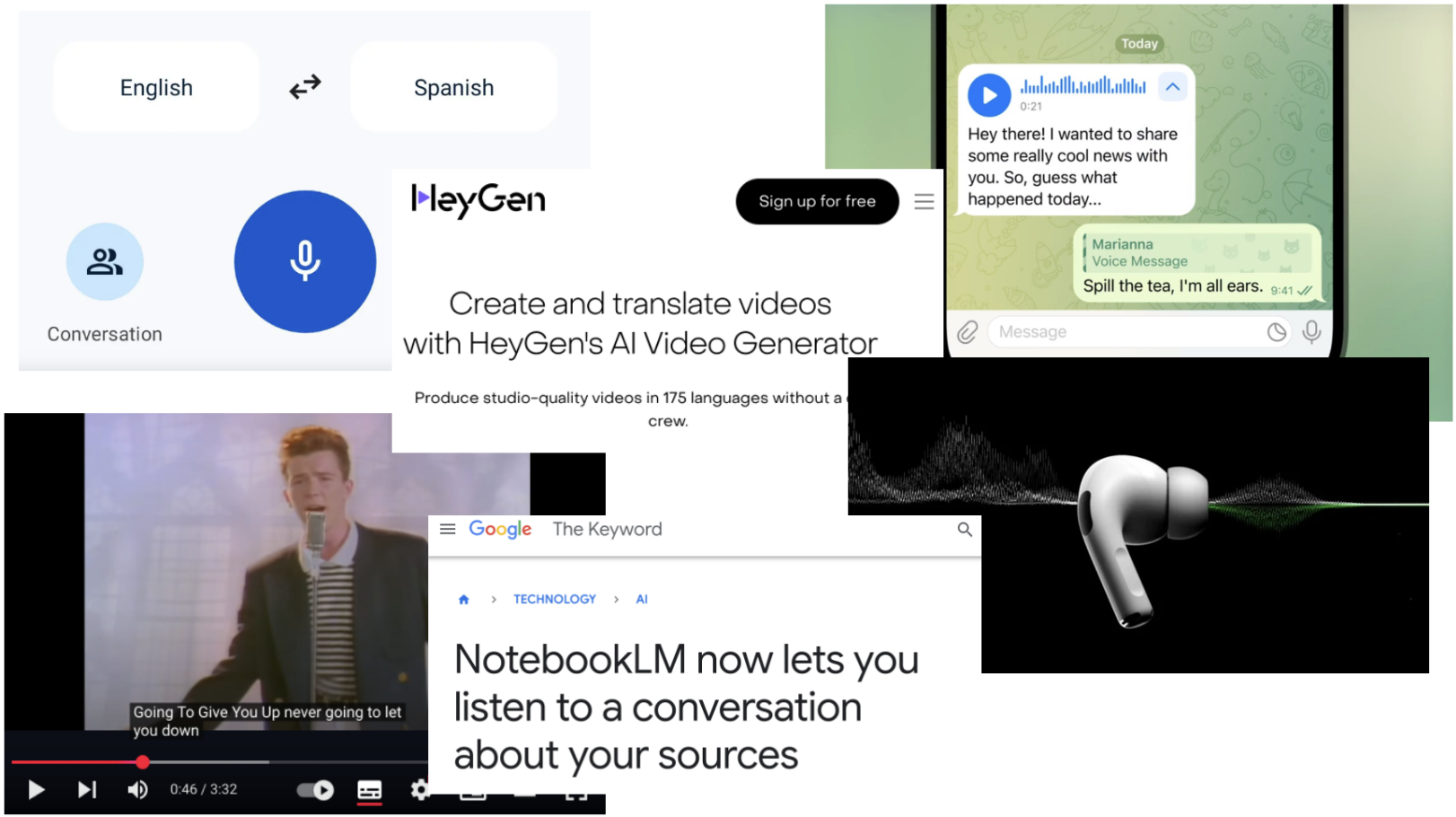
1. Basics:
   - Mathematical analysis, Fourier transform
   - Probability theory
   - Machine learning, Deep learning
2. NLP: self-attention, transformers, tokenization, embeddings
3. Python, PyTorch

Outline

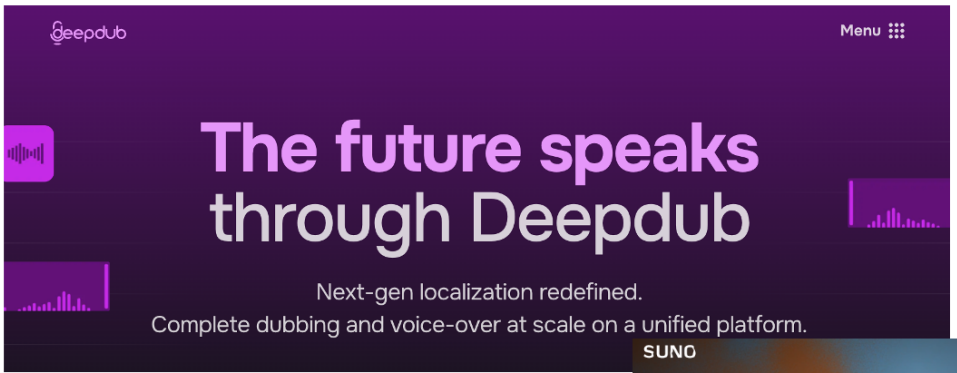
Audio in AI assistants, companions

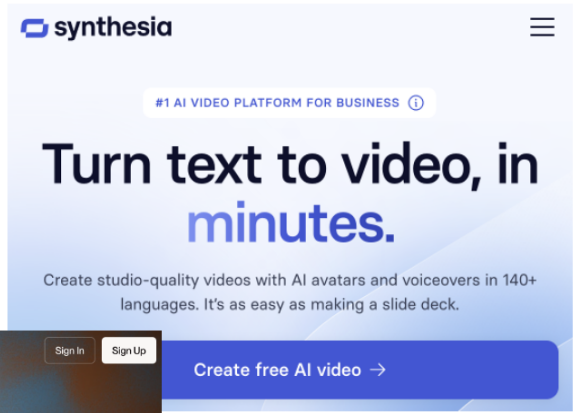


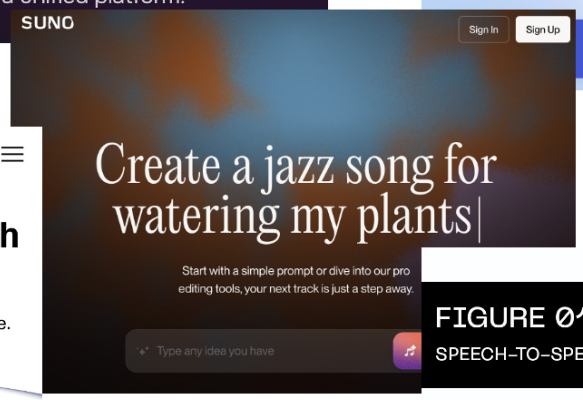
AI & Audio in everyday life

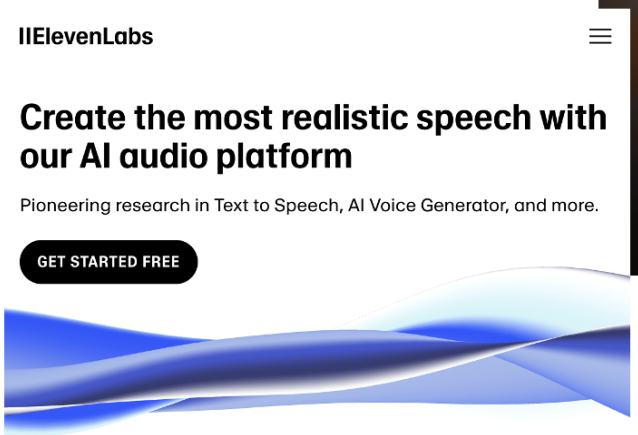


Other Audio & AI applications









Other Audio & AI applications

Meta

Research

Using AI to decode speech from brain activity

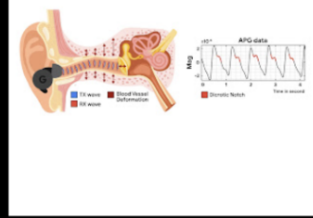
August 31, 2022

Home > Blog >

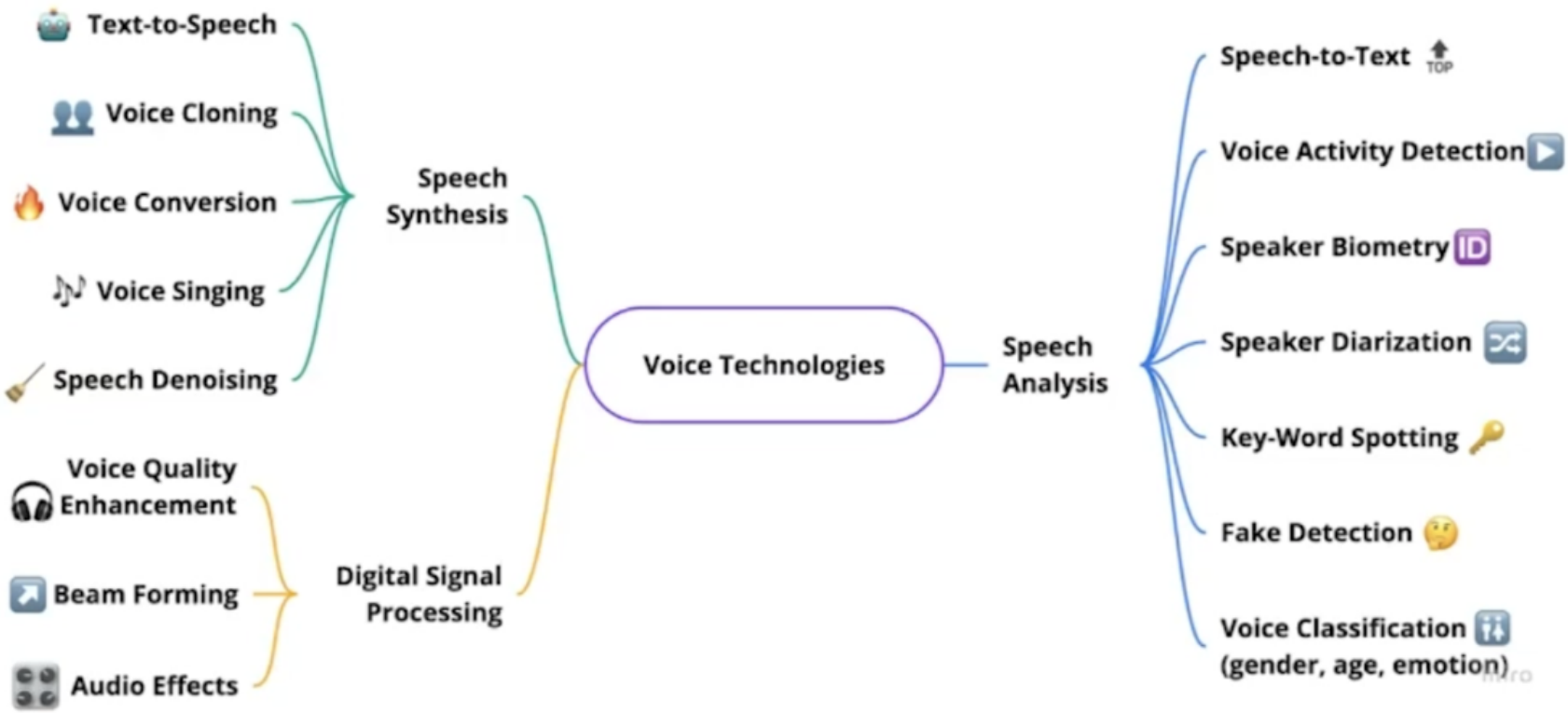
Audioplethysmography for cardiac monitoring with hearable devices

October 27, 2023 ·

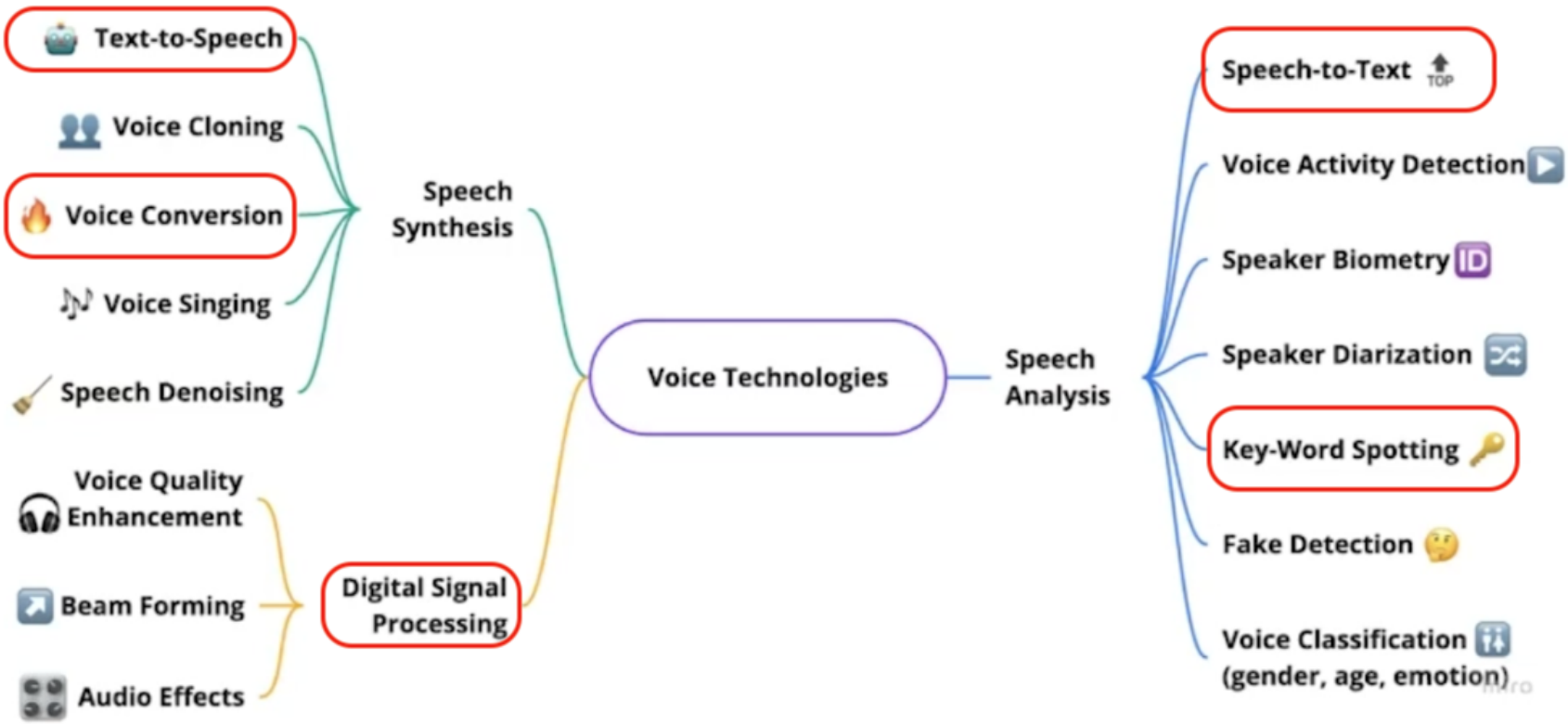
Posted by Xiaoran "Van" Fan, Experimental Scientist, and Trausti Thormundsson, Director, Google



Voice Technologies, Tasks: Mind Map



Voice Technologies, Tasks: Course



Key directions in Audio & AI research

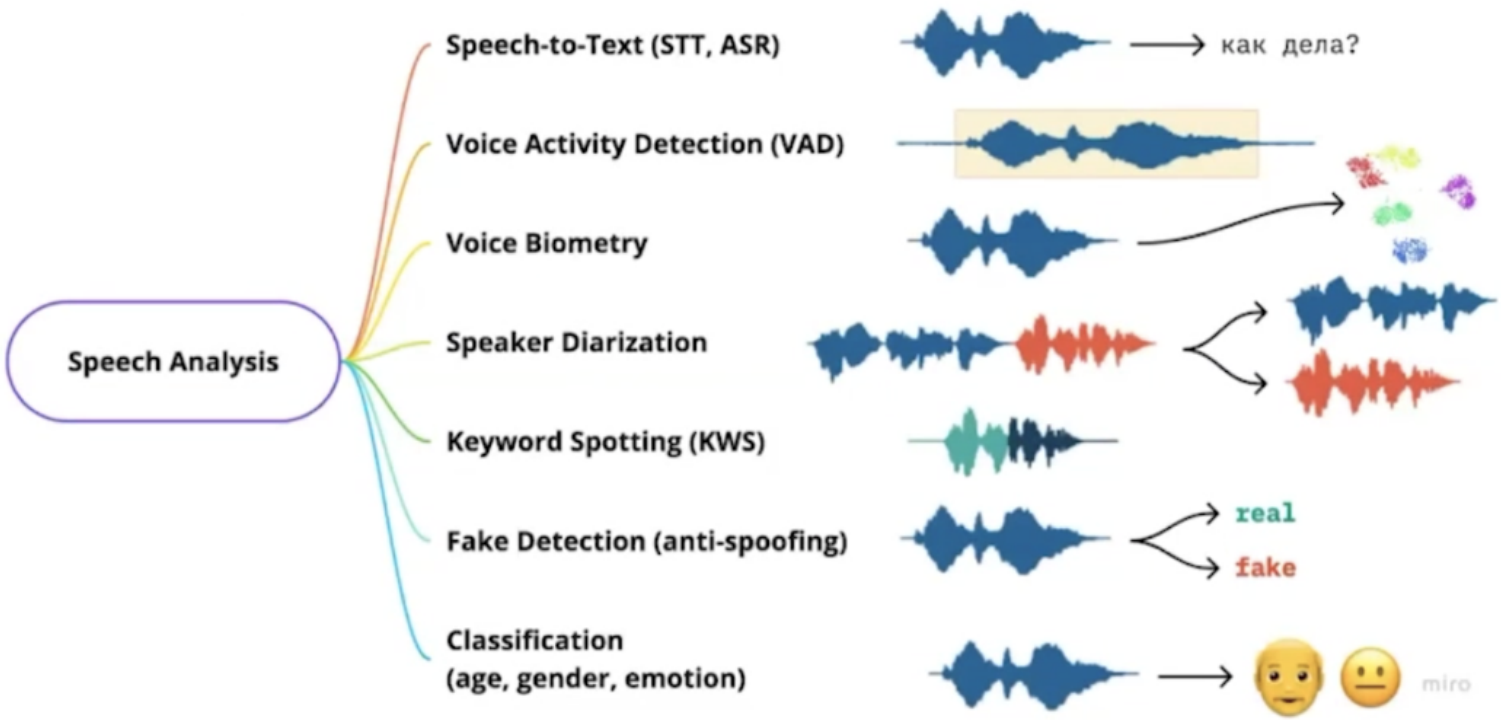
1. Multilingual, Generalized ASR & TTS
2. Audio to audio conversation (End-to-end Differentiable Speech Systems)
3. Multimodal (audio+text+video), ASR/TTS with LLMs
4. Beyond speech: audio & music generation
5. Naturalness, personalization, emotions in TTS
6. Small & fast robust ASR/TTS
7. Real-time Voice Cloning & Speech-to-Speech Translation
8. Spatial Audio Processing
9. DeepFake detection, Security
10. Audio Super Resolution & Restoration

Outline

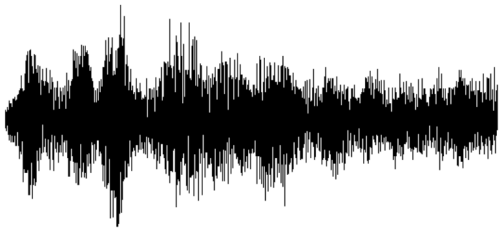
History of Speech Recognition

- **50's**: 1952, Bell Laboratories, "Audrey" system, could recognize single voice speking digits
- **60's**: 1961, IBM, "Shoebox", understood 16 words in English
- **70's**: DARPA, understood over 1000 words (Siri spin-out)
- **80's**: using HMM, understood several thousand words
- **90's**: became faster because of processors
- **00's-10's**: ML, DL, Big Data, GPUs

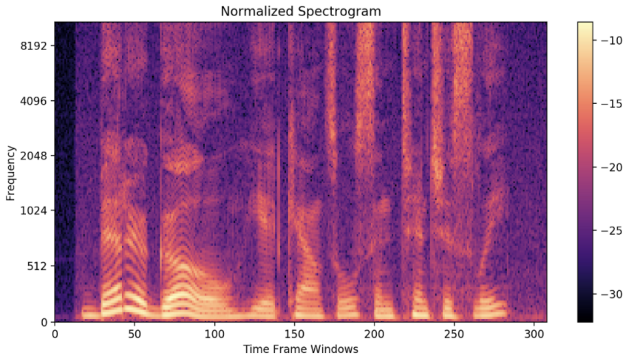
Speech Analysis Tasks: Mind Map



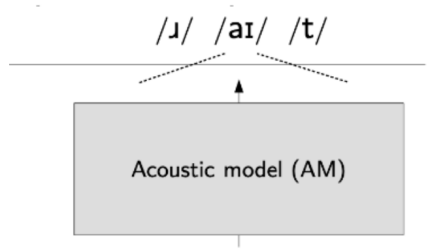
Speech Recognition & Deep Learning: Idea



Raw Audio



Feature: Spectrogram



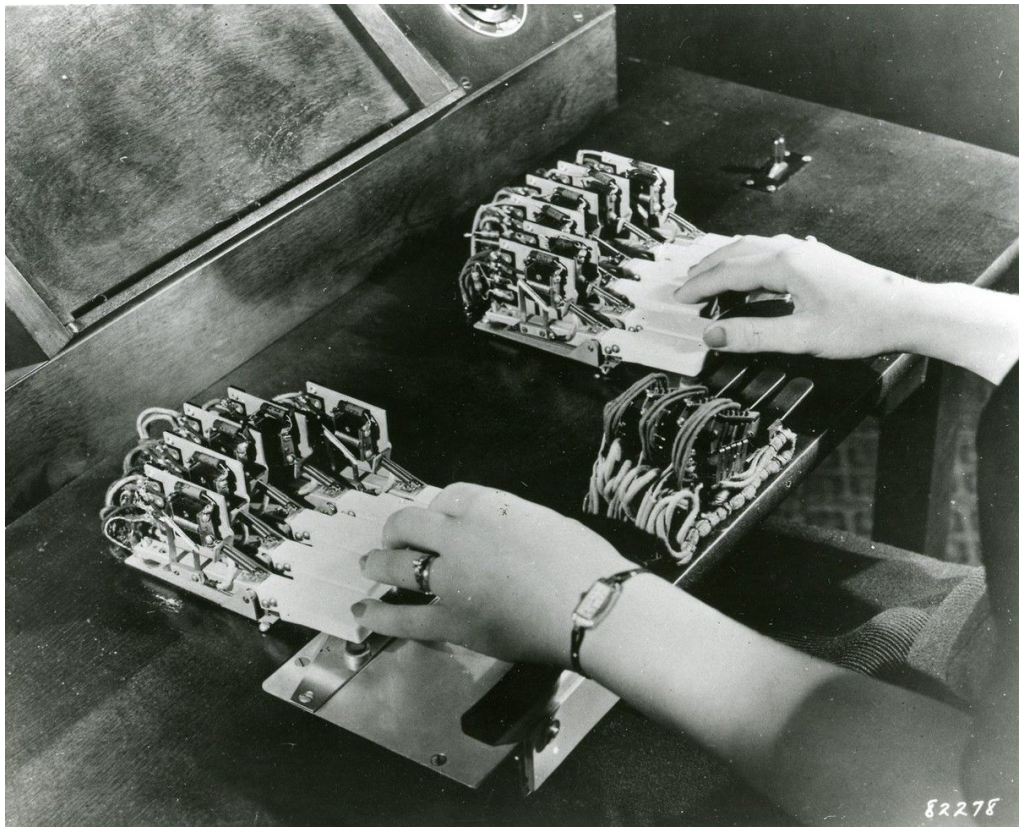
Acoustic model: phonemas

"right"

Language Model

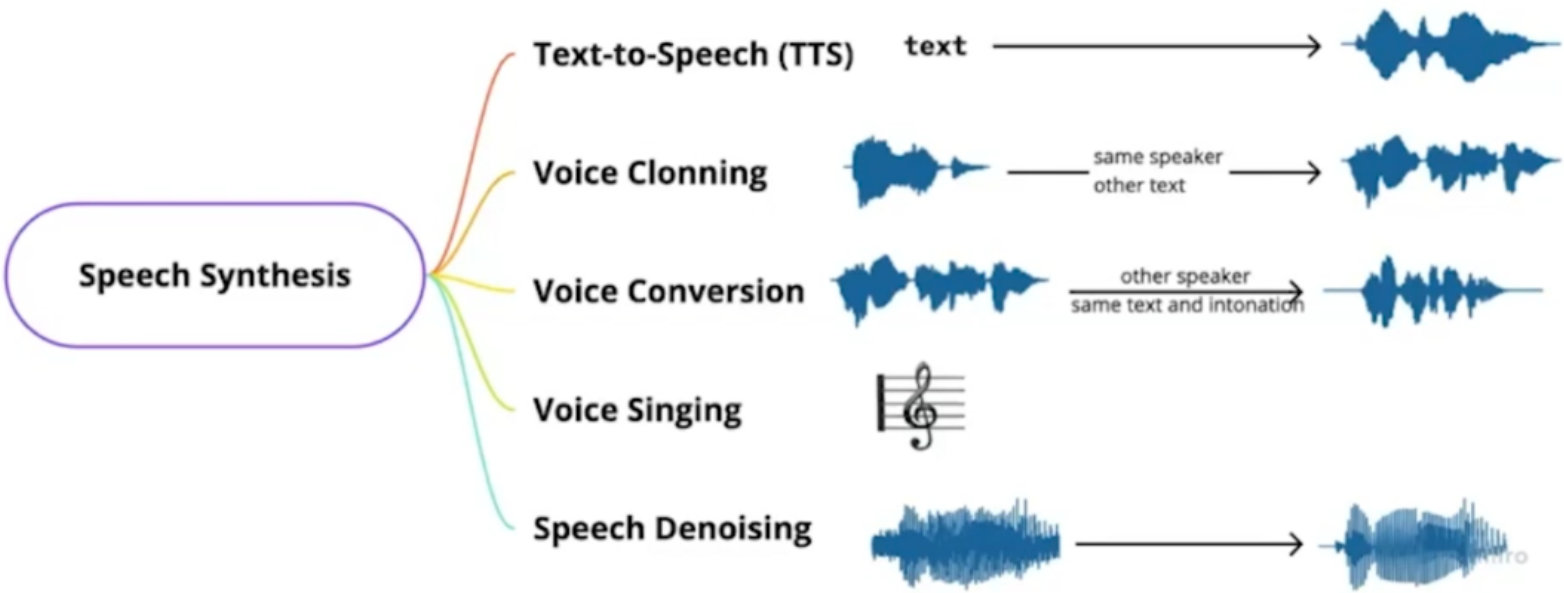
Outline

History of Speech Synthesis



- **30's**: 1939, Bell Laboratories, "Voder",
- **80's**: Format-based on rules, Atari/Sega
- **90's-00's**: Concatenative synthesis
- **10's**: ML, DL, Big Data, GPUs

Speech Synthesis Tasks: Mind Map



Speech Synthesis & Deep Learning: Idea

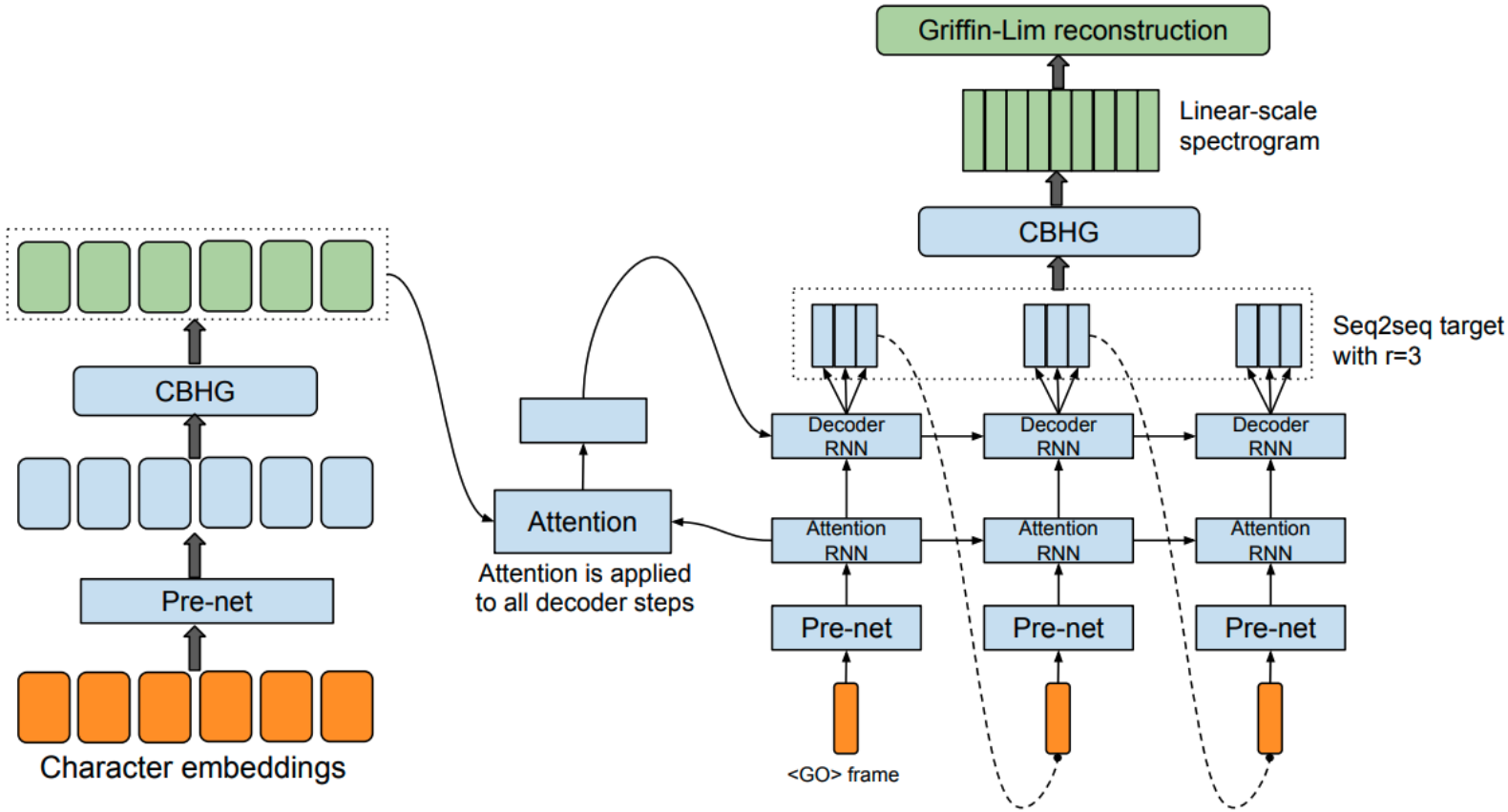


Figure: Example of Deep Learning approach to speech synthesis: encoder-decoder structure with reccurent parts

---

Wang, Yuxuan et al. "Tacotron: Towards End-to-End Speech Synthesis." INTERSPEECH (2017), Google Inc.

Outline

Mind experiment: what is sound?

Figure: If a tree falls in the forest, and there's nobody around to hear, does it make a sound?

What is sound?

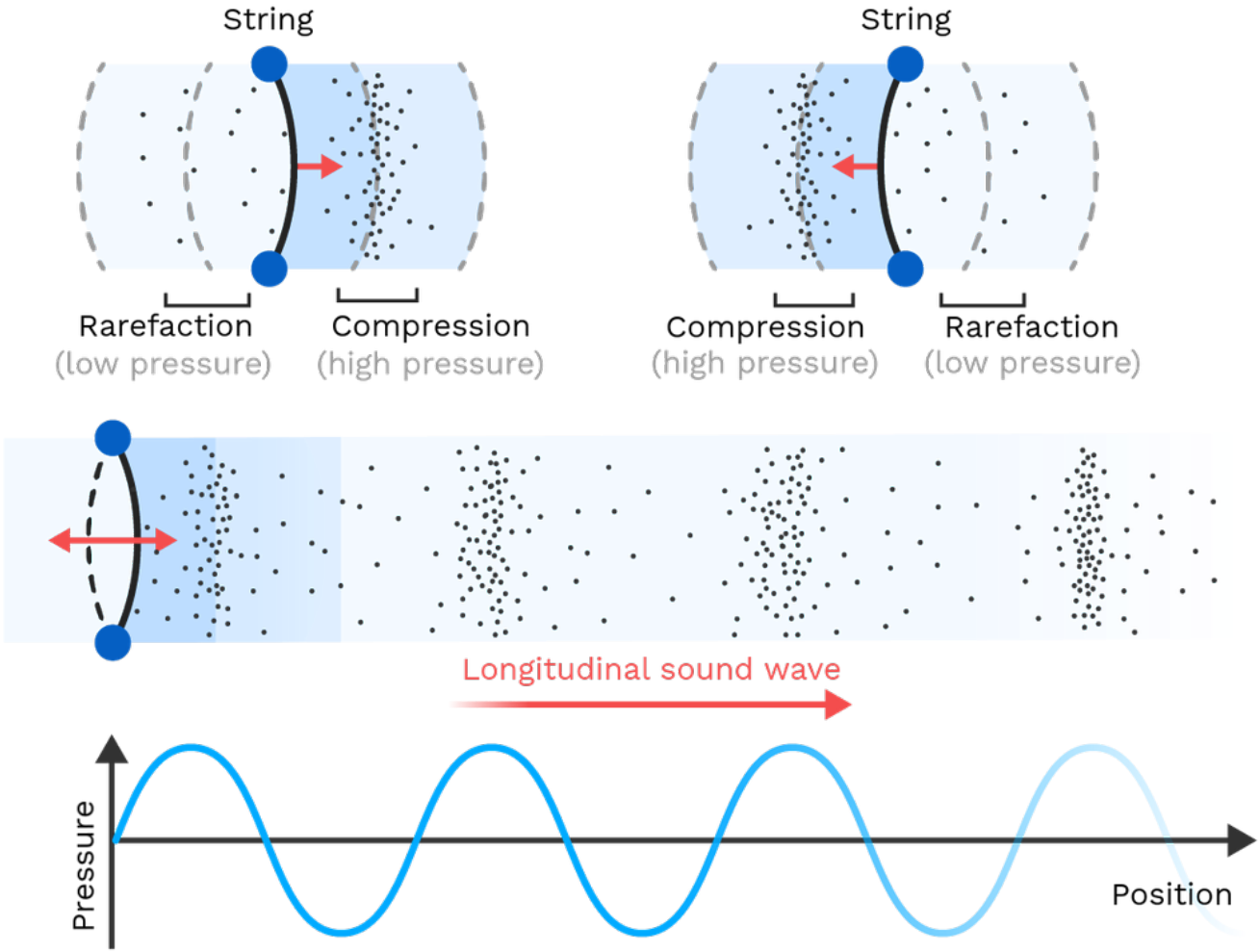
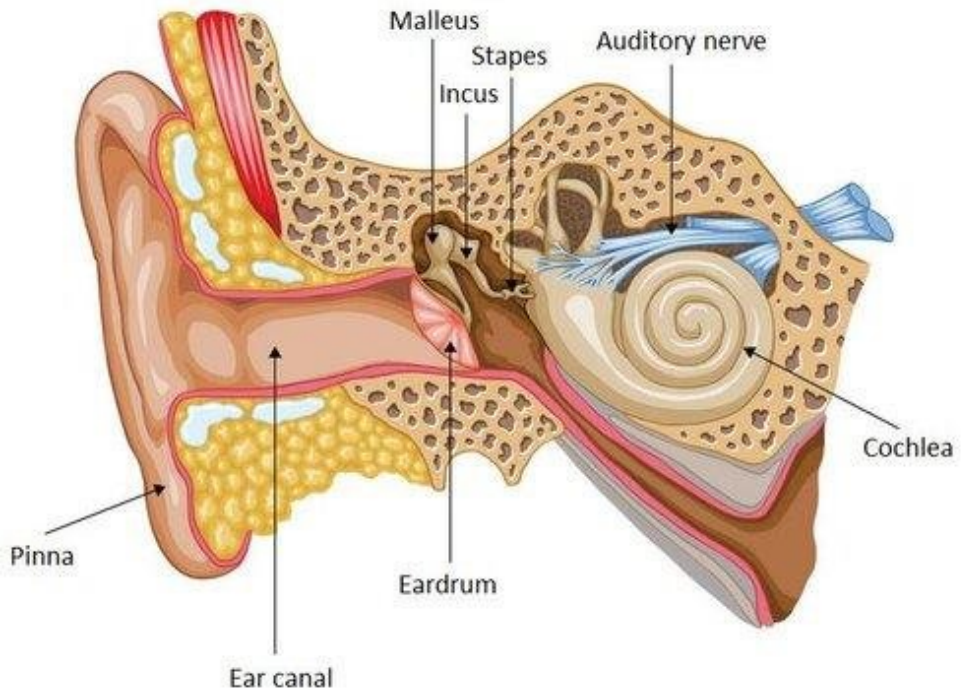


Figure: When we plot pressure, relative to atmospheric pressure, we see the familiar sinusoidal waveform

Ear structure



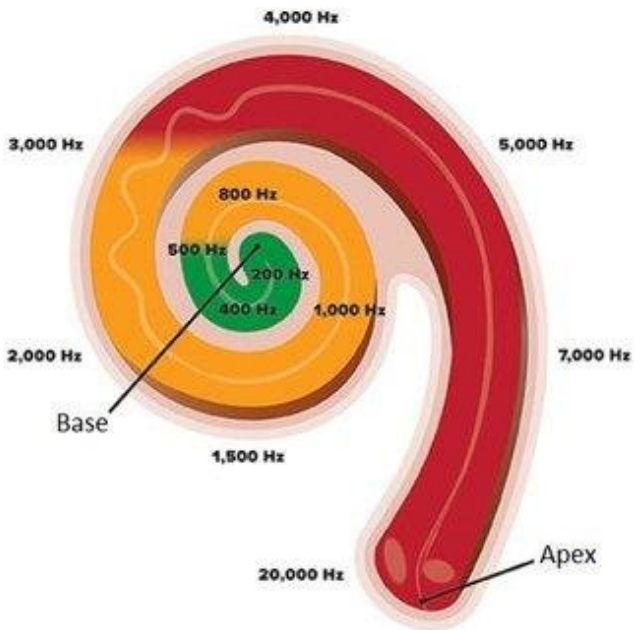


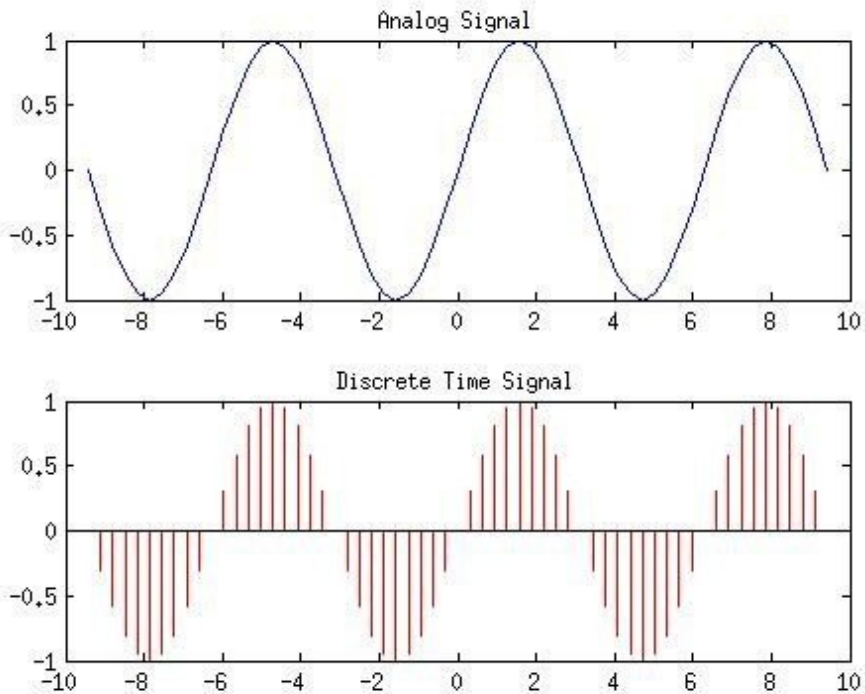
Figure: Fourier decomposition in our body

Fourier transform of a signal $x(t)$

$$\mathcal{F}\{x(t)\} = X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt$$

Sound types: analog vs digital

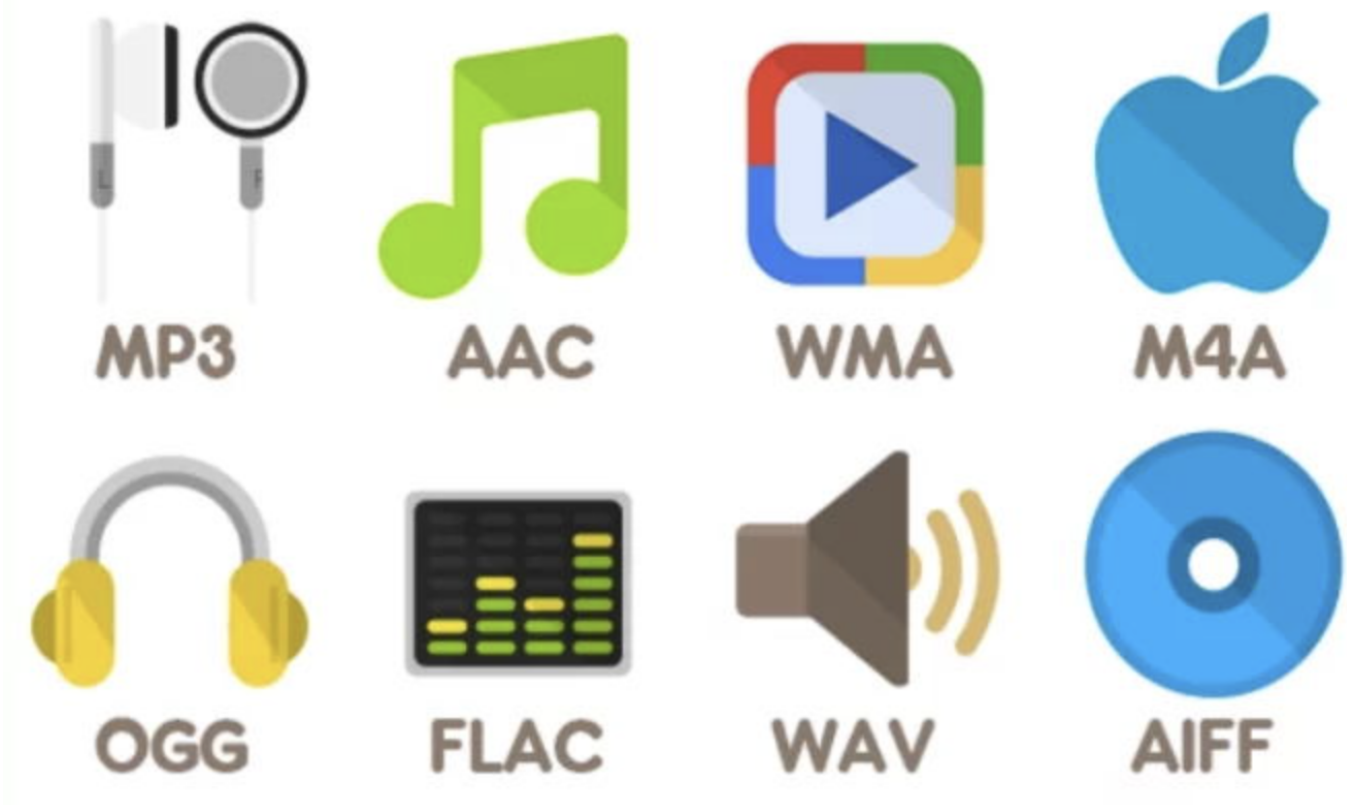
- For any $T > 0$, we may sample a **Continuous-Time** (CT) signal $x(t)$ to generate the **Discrete-Time** (DT) signal $x[n] = x(nT)$
- DT: signal $x(t)$ is evaluating at uniformly spaced points on the t-axis. The number T is the **sampling period**
- **Sampling frequency**: $f_s = \frac{1}{T}$, [Hertz or samples/sec.]



Sound characteristics

- Signal $x(t)$
- Signal energy $\int_{-\infty}^{\infty} |x(t)|^2 dt$ (used for normalizing signals, augmentations)
- Sample rate (SR) – number of audio samples per one second
  - 8 kHz or 16 kHz - standart for audio in telephony
  - 44.1 kHz - CD audio/Computer Audio
  - 48 kHz - DVD audio/Computer Audio
  - 96 kHz - High resolution Audio
- Number of channels – how many signals we record in parallel (mono: 1, stereo: 2)

Audio file formats



- Uncompressed: WAV, AIFF
- Lossless compression: FLAC, ALAC
- Lossy compression: MP3, Opus, AAC, OGG, WMA

The Nyquist Theorem

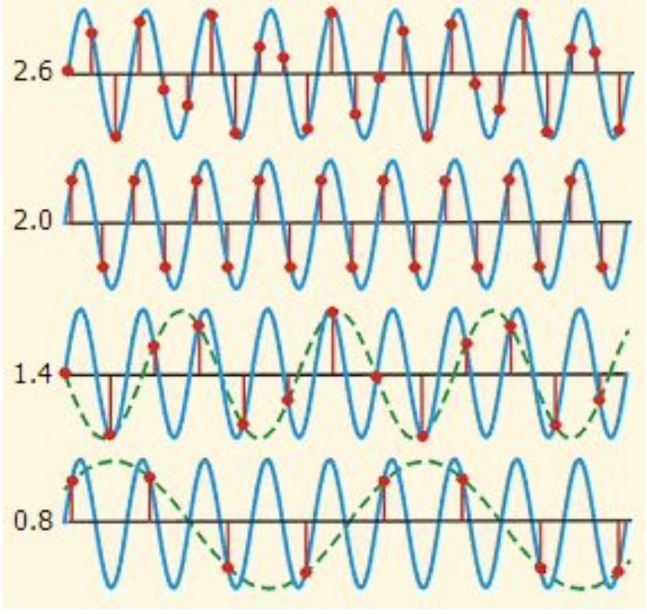
- Σ_{CT} – set of all CT signals $x(t)$, Σ_{DT} – set of all DT signals $x[n]$. Procedure of sampling for a given sampling period T:

$$\Sigma_{CT} \longmapsto \Sigma_{DT} \quad \Leftrightarrow \quad x(t) \longmapsto x[n] = x(nT)$$

- CT signal $x(t)$ is bandlimited if there exists $\omega_B < \infty$ such that $X_{CT}(j\omega) = 0$ for $|\omega| > \omega_B$

Nyquist Theorem

The sampling map is bijection on Σ_{ω_B} iff $\omega_s > 2\omega_B$.
$\omega_s = 2\pi f_s = \frac{2\pi}{T}$ – radian sampling frequency



Outline

Fourier Transform: motivation

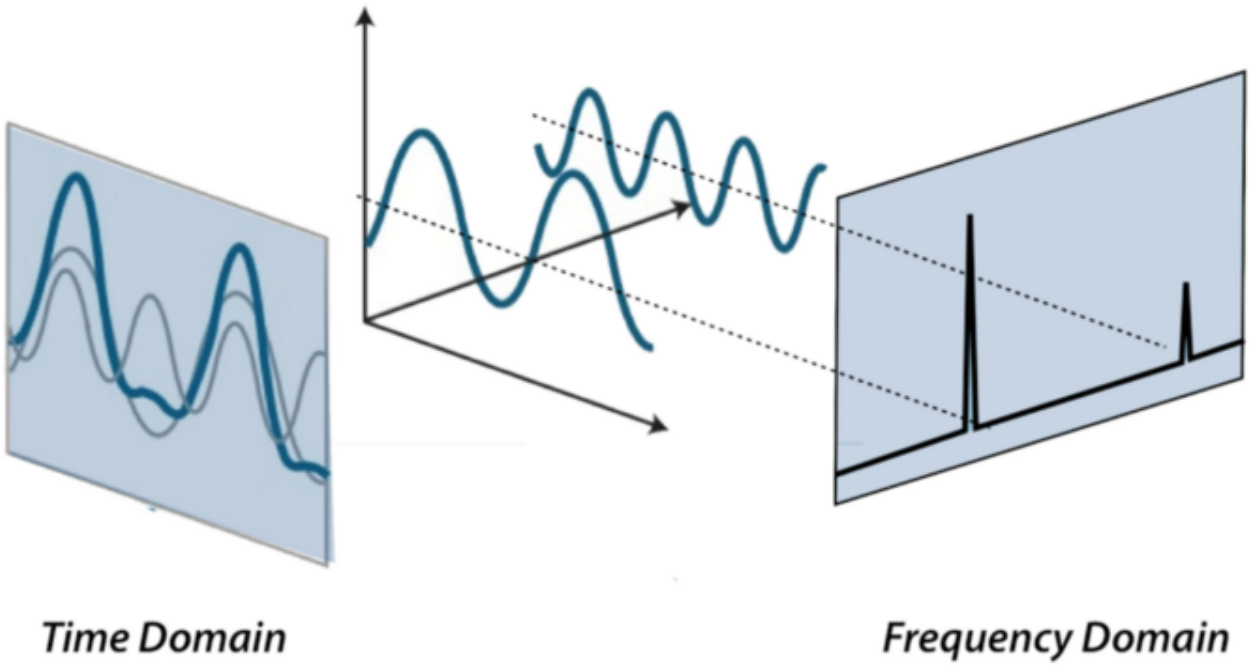


Figure: Fourier Transform transfer a signal from the **time domain** to the **frequency domain**

Fourier Transform

- CT Fourier transform (CTFT) of a CT signal $x(t)$ is

$$\mathcal{F}\{x(t)\} = X(j\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt$$

- CT unit impulse $\delta(t)$, $\int_{-\infty}^{\infty} \delta(t)dt = 1$,
- Define signal by CT impulse $x(t) = \sum_{n=-\infty}^{\infty} x[n]\delta(t-n)$
- Taking Fourier transform for DT signal:

$$\begin{aligned} X(j\omega) &= \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} x[n]\delta(t-n)e^{-j\omega t}dt \\ &= \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n} \int_{-\infty}^{\infty} \delta(t-n)e^{-j\omega(t-n)}dt \\ &= \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n} = \mathbf{M}x(t). \end{aligned}$$

DTFT: example

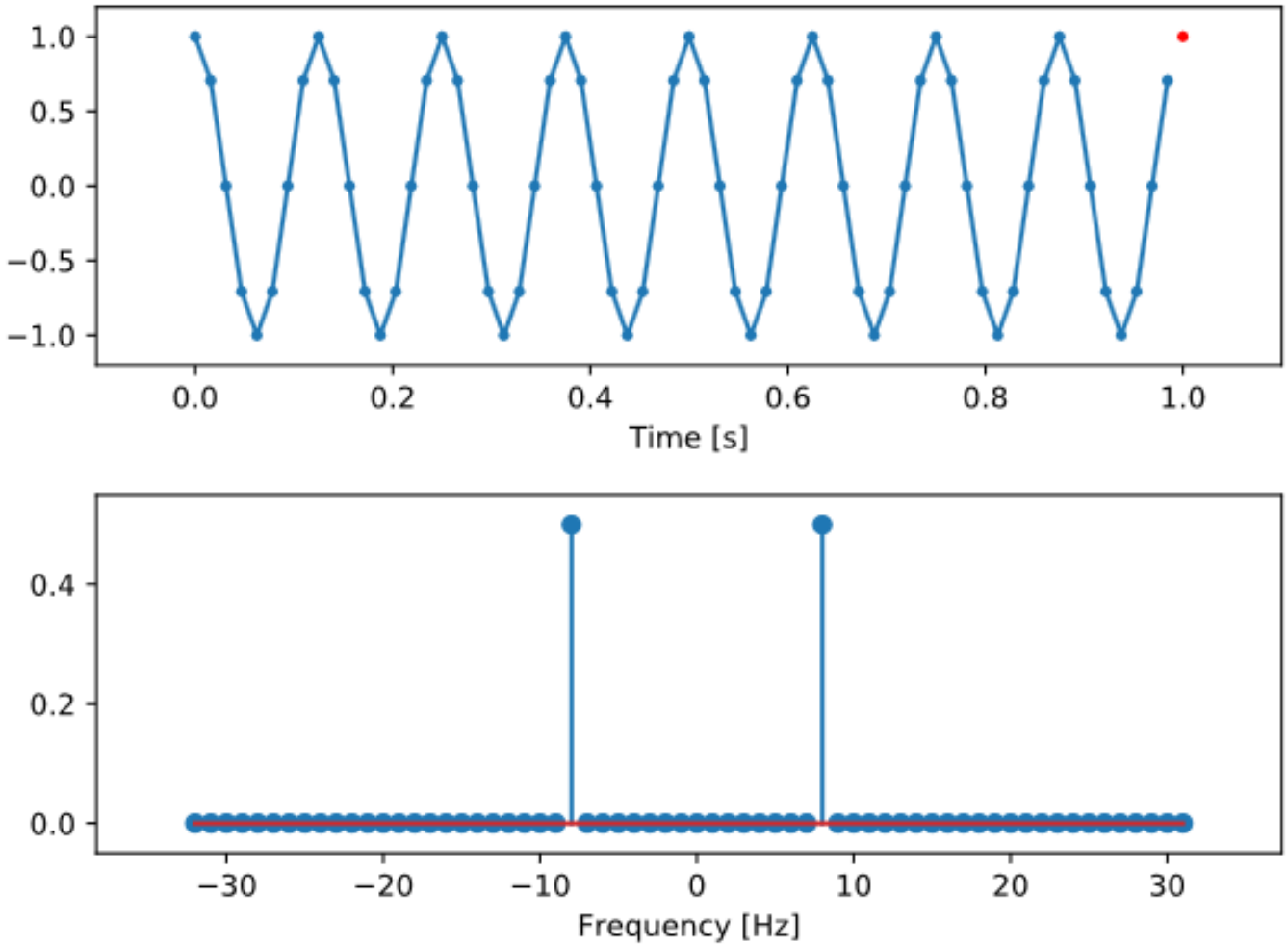
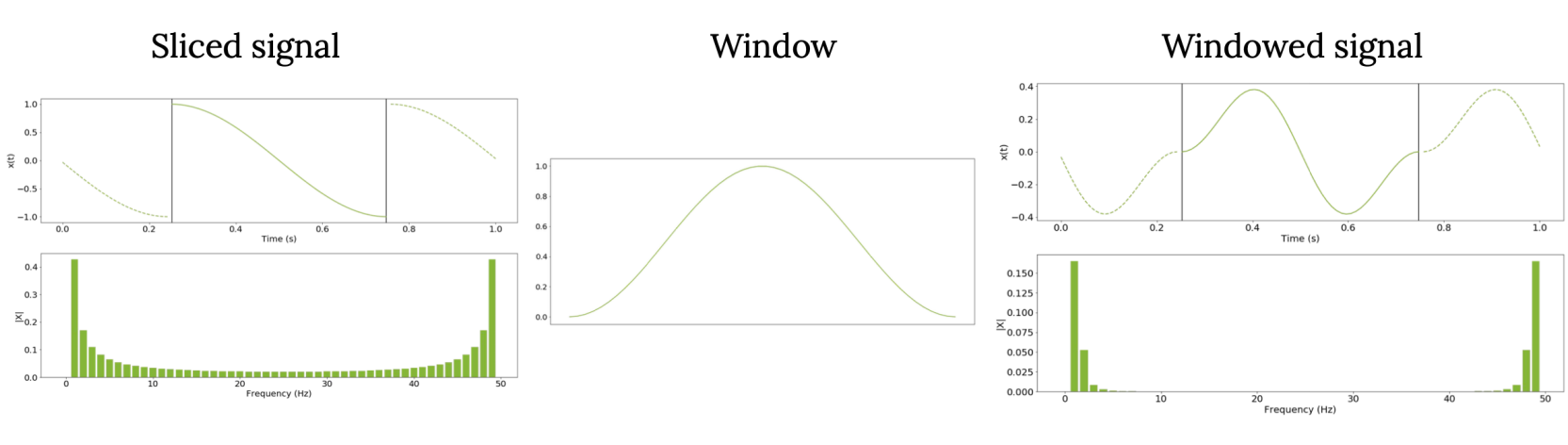
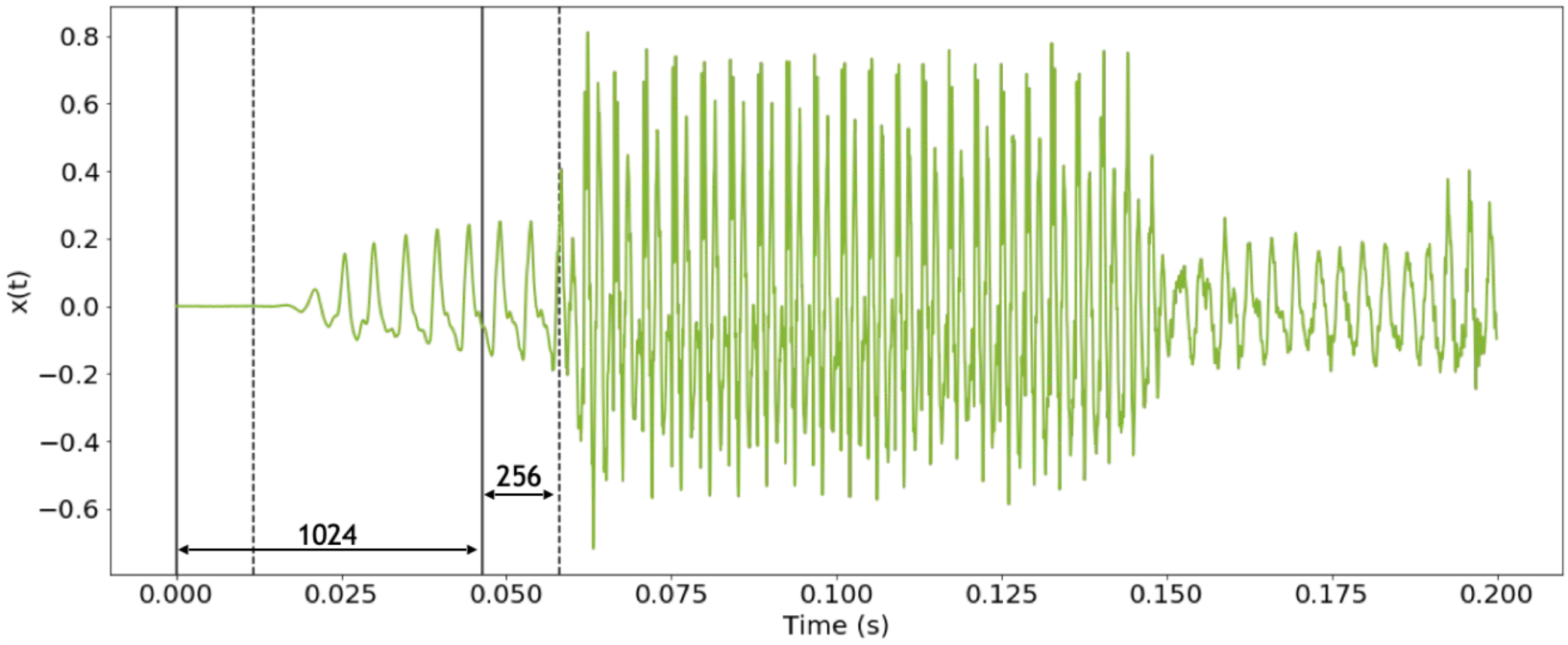


Figure: Example of DTFT for cosine signal

Problems with Fourier Transorm in real life



Short-time Fourier Transform (STFT)



STFT (DTFT with Hamming window):

$$X[r, w) = \sum_{n=-\infty}^{\infty} w[r - n]x[n]e^{-j\omega n},$$

where w – window function, r – location of window along the time axis

Outline

Spectrograms

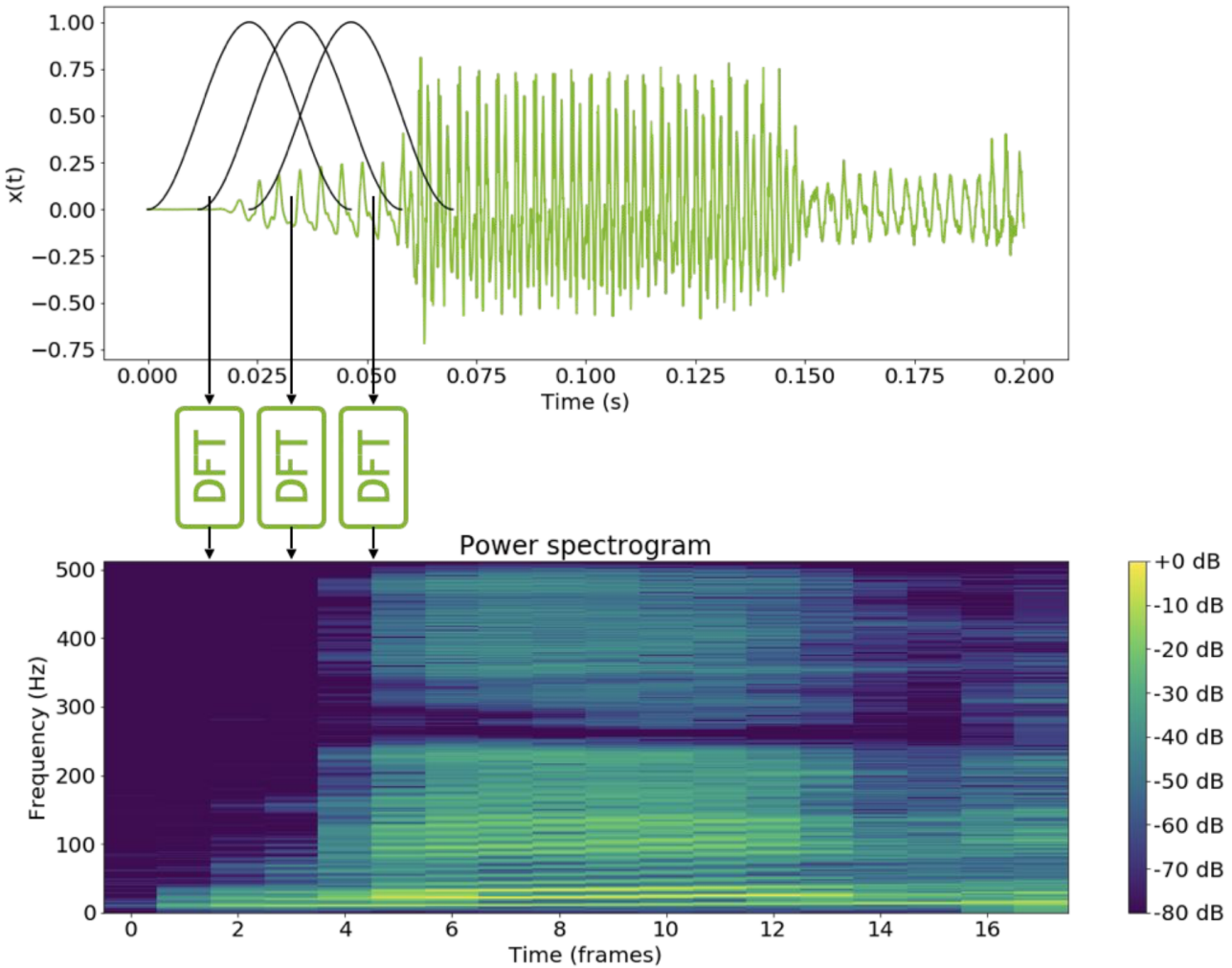
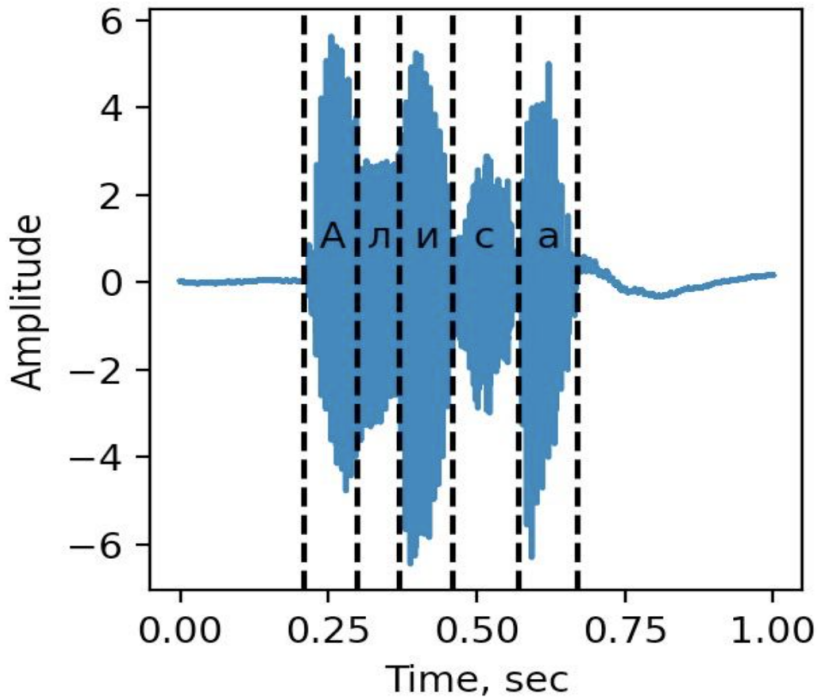
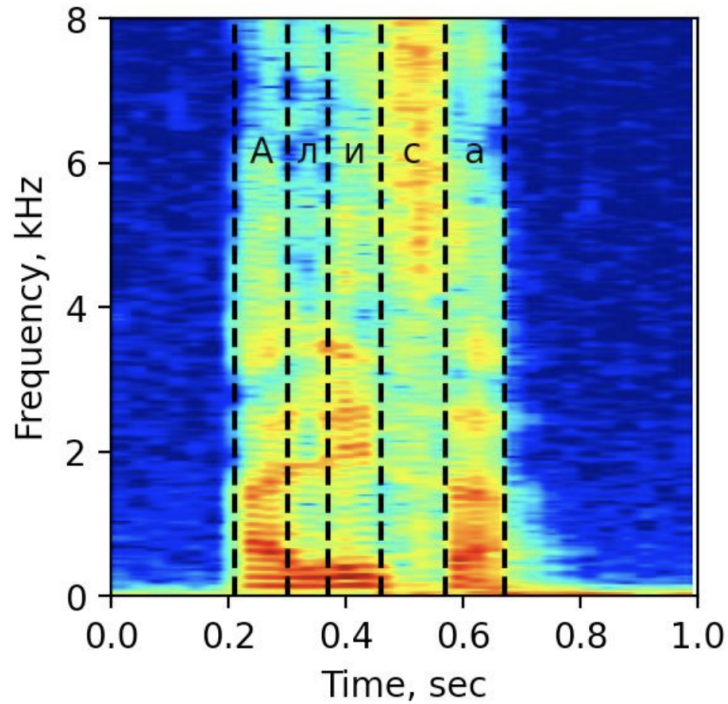


Figure: STFT+window Spectrogram

Example: Why to use spectrograms





Mel Scale

- Mel Scale: humans percieve sound on a log-scale, not linear
- $500Hz \ll 600Hz$, but $5000Hz \approx 5100Hz$
- A lot of formulas to convert f hertz into m mels. Popular example:

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

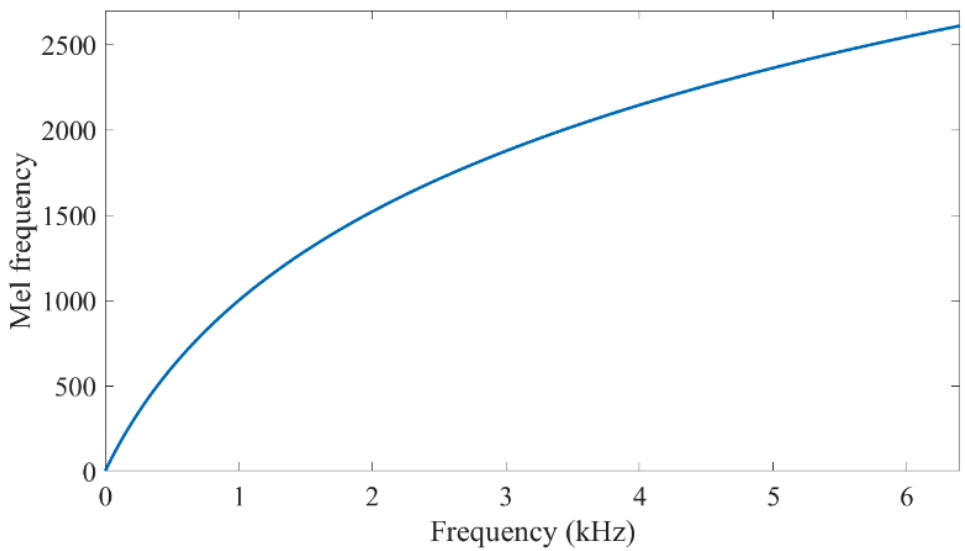
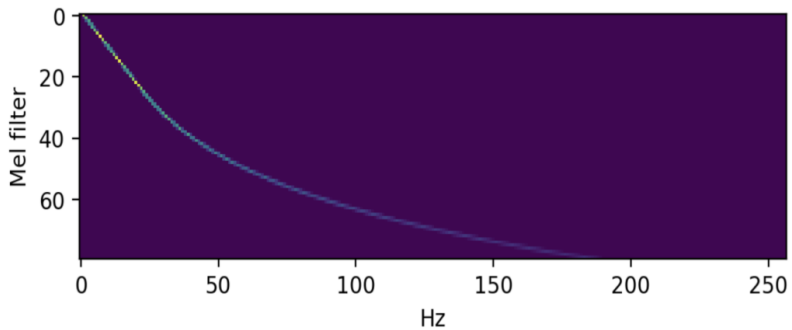
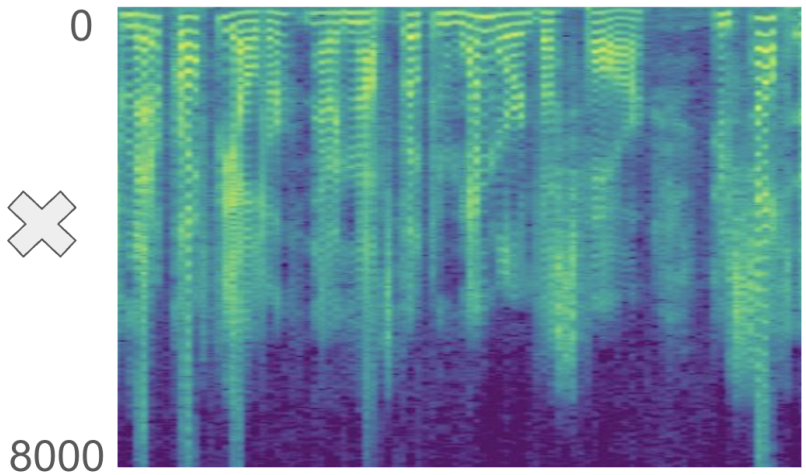


Figure: Mel scale

Mel Filter Bank

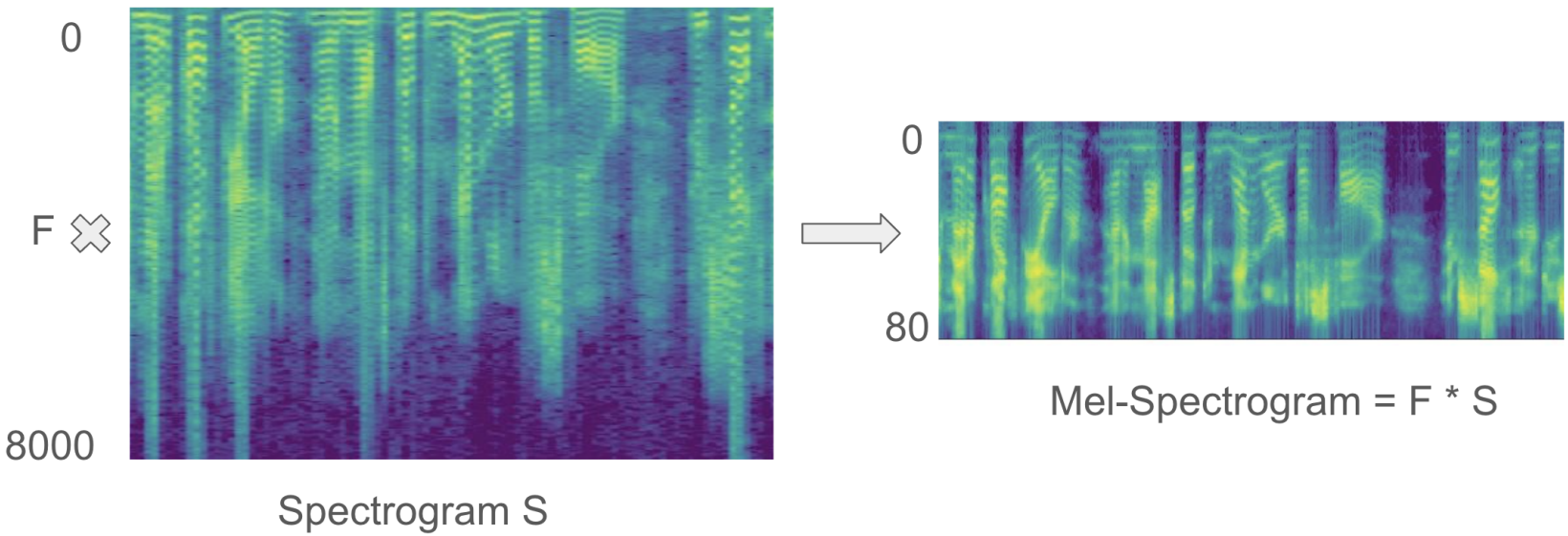


Filterbank F



Spectrogram S

Mel Filter Bank



Fundamental frequency (F_0)

- Sound $\approx$ Fundamental frequency + resonances + harmonics
- Fundamental frequency F_0 is the physical source frequency (Pitch $\approx F_0$, Pitch is perceptual value, F_0 - physical)
- Harmonics $= k * F_0$ (determine the timbre)
- F_0, harmonics depend on the vocal cords
- Formants ($F_1, F_2 ...$): determined by a tongue, lips and oral cavity

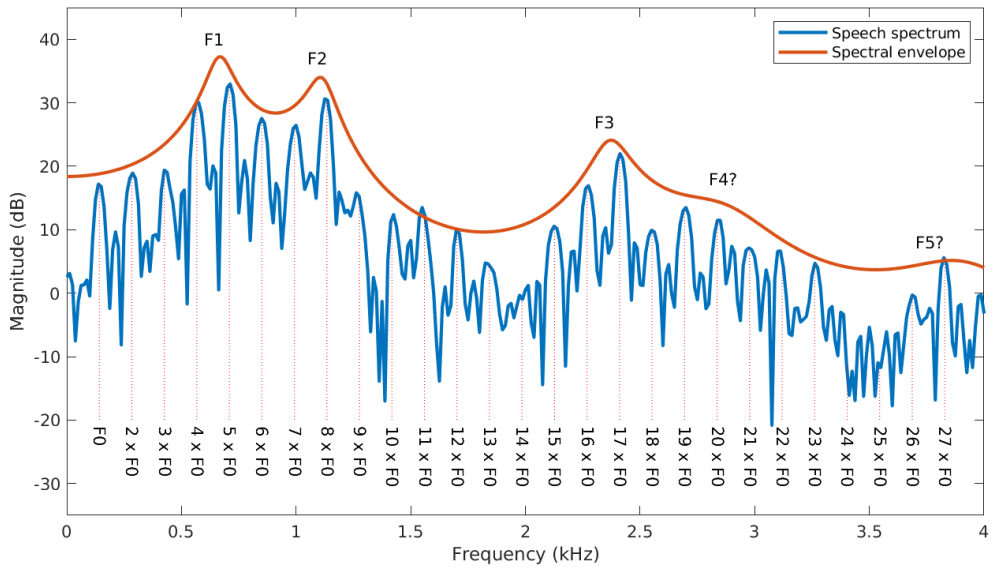
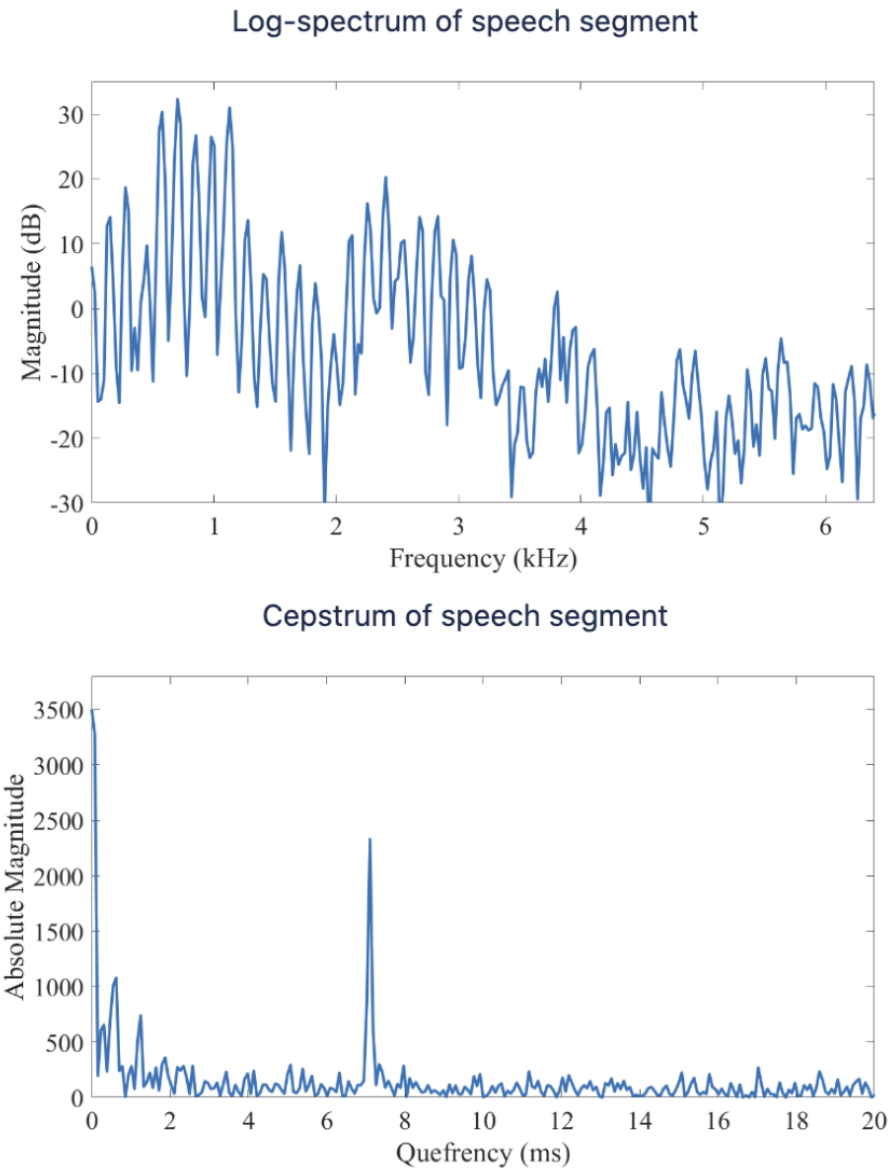


Figure: Spectral envelope curve: peaks are formants

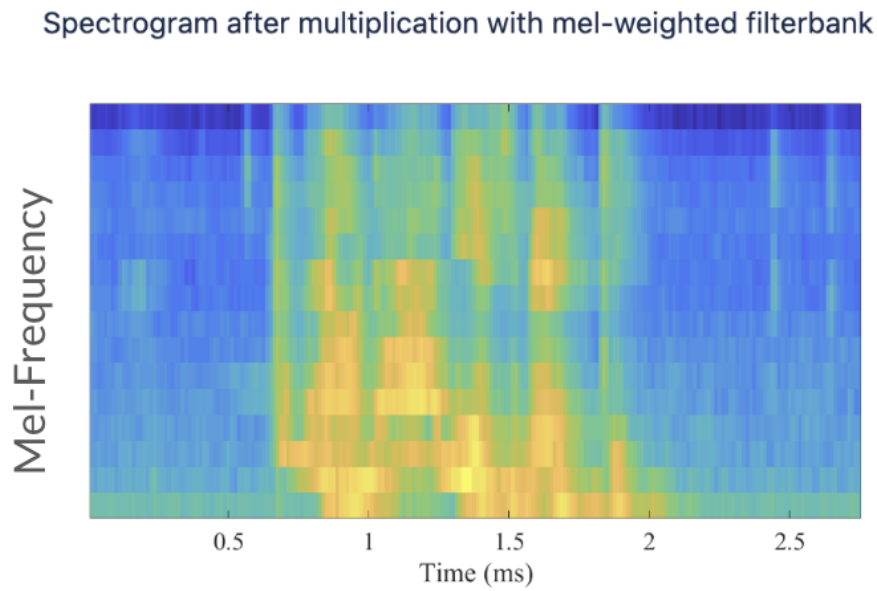
Cepstrum

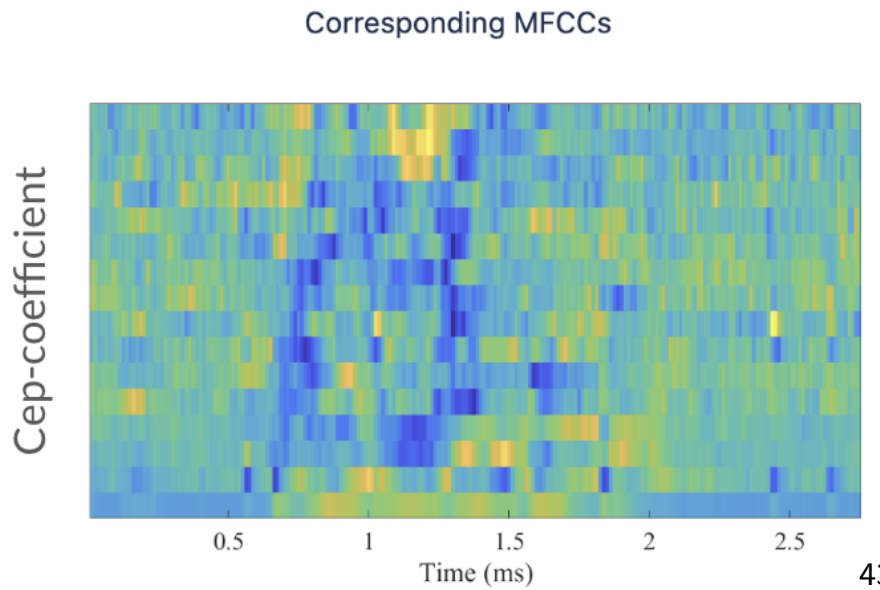
- Fourier spectrum of voice has **periodic** structure
- Apply **Inverse DFT** to log-spectrum and obtain **Cepstrum**
- **Peak** in Cepstrum should be located at $\frac{1}{F_0}$



MFCC (Mel-Frequency Cepstral Coefficients)

1. Apply STFT to the signal
2. Apply mel filters
3. Take the log value
4. Apply Discrete Cosine Transform





Spectrogram: summary

- Waveform
- **Spectrogram**: STFT+window
- **Mel Spectrogram**: STFT+window+Mel scale
- **MFCC**: STFT+window+Mel scale+log+DCT