Deep Learning Audio

Lecture 2

Pavel Severilov

Moscow Institute of Physics and Technology

2022

Outline

1. Sound: Introduction
2. Sound: characteristics
3. Fourier Transform
4. Spectrograms

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?

Ear structure

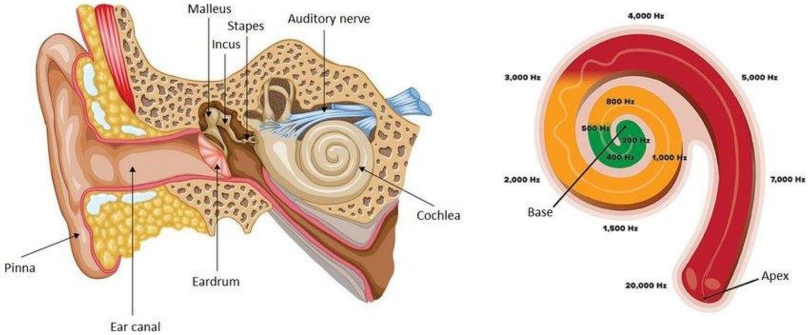


Figure: Fourier decomposition in our body

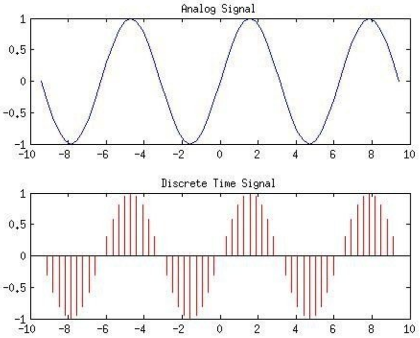
Fourier transform of a signal $x(t)$

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

Outline

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)

The Nyquist Theorem

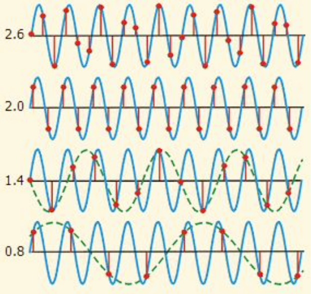
- Σ_{CT} – set of all CT signals $x(t)$, Σ_{DT} – set of all DT signals $x[n]$. Procedure of sampling for a given sampling period T:

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

- 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

1. Sound: Introduction
2. Sound: characteristics
3. Fourier Transform
4. Spectrograms

Fourier Transform: motivation

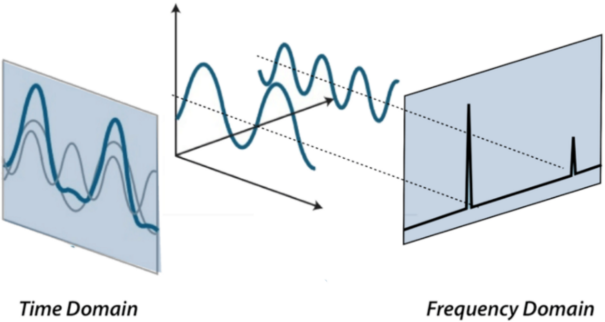


Figure: Fourier Transform transfer a signal from the **time domain** to the **frequency domain**

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

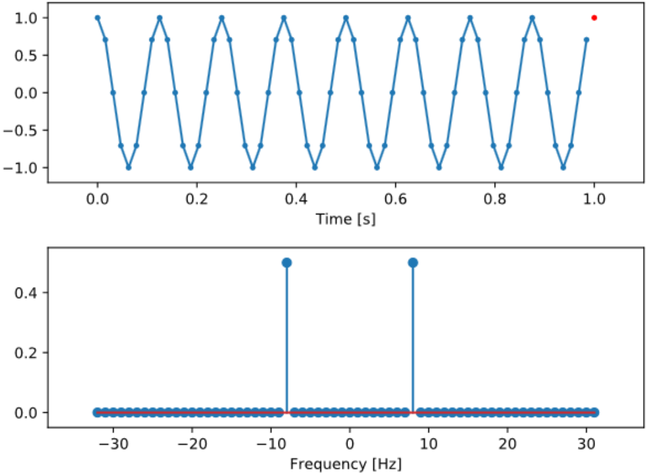
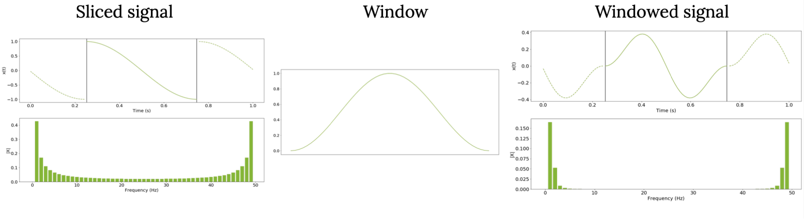
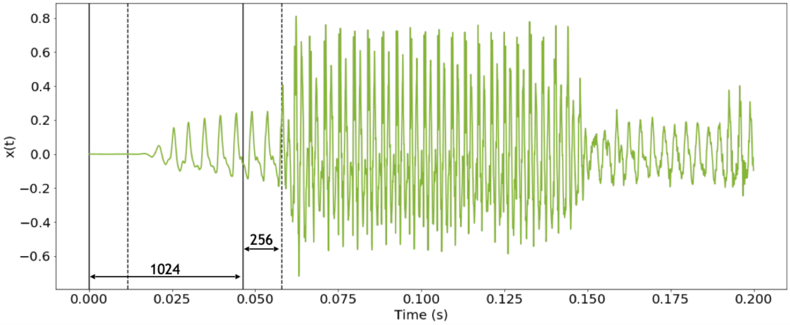


Figure: Example of DTFT for cosine signal

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

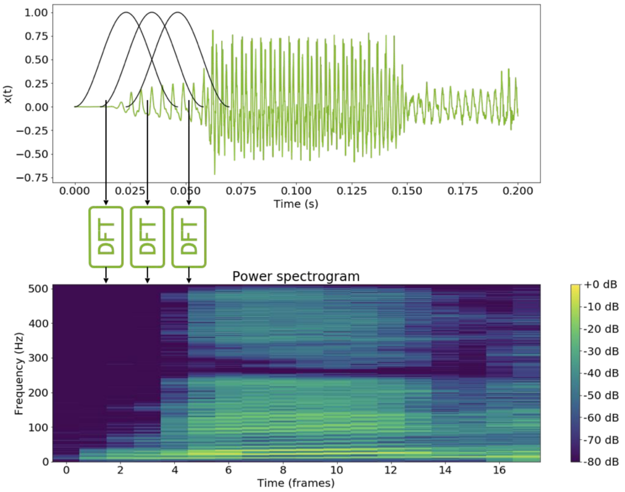


Figure: STFT+window Spectrogram

Mel Scale

- Mel Scale: humans percieve sound on a log-scale, not linear
- A lot of formulas to convert f hertz into m mels. Popular example:

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

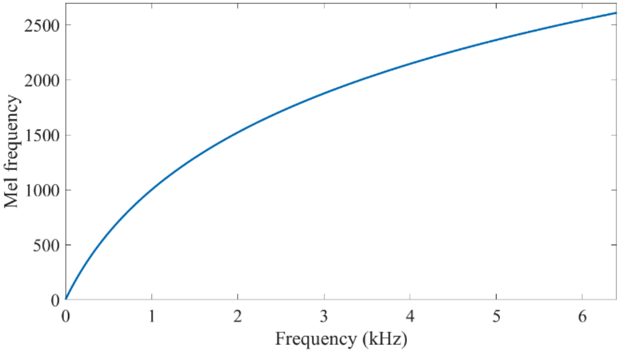


Figure: Mel scale

Mel spectrogram

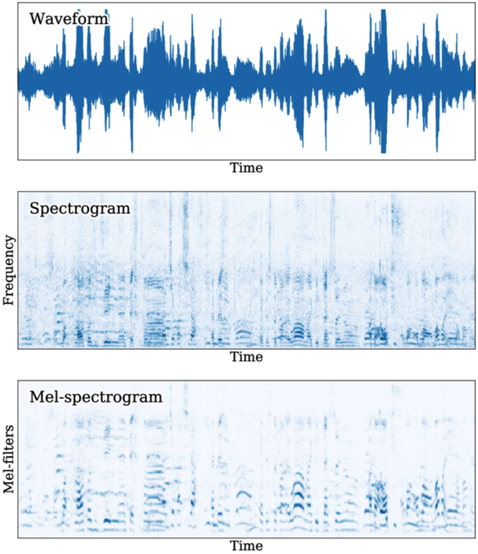


Figure: Waveform $\rightarrow$ STFT+window Spectrogram $\rightarrow$ STFT+window+Mel scale Spectrogram