# 2 STFT

The Fourier transform and in particular the discrete STFT serve as *front-end transform*, the first computing step, for deriving a large number of different musically relevant audio features. We now recall the definition of the discrete STFT while fixing some notation. Let $x : [0 : L-1] := \{0, 1, \ldots, L-1\} \to \mathbb{R}$ be a real-valued discrete-time signal of length $L$ obtained by equidistant sampling with respect to a fixed sampling rate $F_\mathrm{s}$ given in Hertz (Hz). Furthermore, let $w : [0 : N-1] := \{0, 1, \ldots, N-1\} \to \mathbb{R}$ be a discrete-time window of length $N \in \mathbb{N}$ (usually a power of two) and let $H \in \mathbb{N}$ be a hop size parameter. With regards to these parameters, the discrete STFT $\mathcal{X}$ of the signal $x$ is given by

$$\mathcal{X}(m, k) := \sum_{n=0}^{N-1} x(n + mH) w(n) \exp(-2\pi i k n / N) \tag{1}$$

with $m \in [0 : \lfloor \frac{L-N}{H} \rfloor]$ and $k \in [0 : K]$. The complex number $\mathcal{X}(m, k)$ denotes the $k^{\mathrm{th}}$ Fourier coefficient for the $m^{\mathrm{th}}$ time frame, where $K = N/2$ is the frequency index corresponding to the Nyquist frequency. Each Fourier coefficient $\mathcal{X}(m, k)$ is associated with the physical time position (using the start position of the window as reference point)

$$T_{\mathrm{coef}}(m) := \frac{m \cdot H}{F_\mathrm{s}} \tag{2}$$

given in seconds (sec) and with the physical frequency

$$F_{\mathrm{coef}}(k) := \frac{k \cdot F_\mathrm{s}}{N} \tag{3}$$

given in Hertz (Hz). For example, using $F_\mathrm{s} = 44100$ Hz as for a CD recording, a window length of $N = 4096$, and a hop size of $H = N/2$, we obtain a time resolution of $H/F_\mathrm{s} \approx 46.4$ ms and frequency resolution of $F_\mathrm{s}/N \approx 10.8$ Hz.

**Homework Exercise 1**

(a) Compute the time and frequency resolution of the resulting STFT when using the following parameters. What are the Nyquist frequencies?
   (i) $F_\mathrm{s} = 22050$, $N = 1024$, $H = 512$
   (ii) $F_\mathrm{s} = 48000$, $N = 1024$, $H = 256$
   (iii) $F_\mathrm{s} = 4000$, $N = 4096$, $H = 1024$

(b) Using $F_\mathrm{s} = 44100$, $N = 2048$ and $H = 1024$, what is the physical meaning of the Fourier coefficients $\mathcal{X}(1000, 1000)$, $\mathcal{X}(17, 0)$, and $\mathcal{X}(56, 1024)$?

The STFT is often visualized by means of a *spectrogram*, which is a two-dimensional representation of the squared magnitude:

$$\mathcal{Y}(m, k) = |\mathcal{X}(m, k)|^2. \tag{4}$$

When generating an image of a spectrogram, the horizontal axis represents time, the vertical axis is frequency, and the dimension indicating the spectrogram value of a particular frequency at a particular time is represented by the intensity or color in the image.