ECS7006 Music Informatics

Week 2 - Time-Frequency Representations

School of Electronic Engineering and Computer Science
Queen Mary University of London

prepared by Emmanouil Benetos
adapted from material by Meinard Müller and Simon Dixon

emmanouil.benetos@qmul.ac.uk

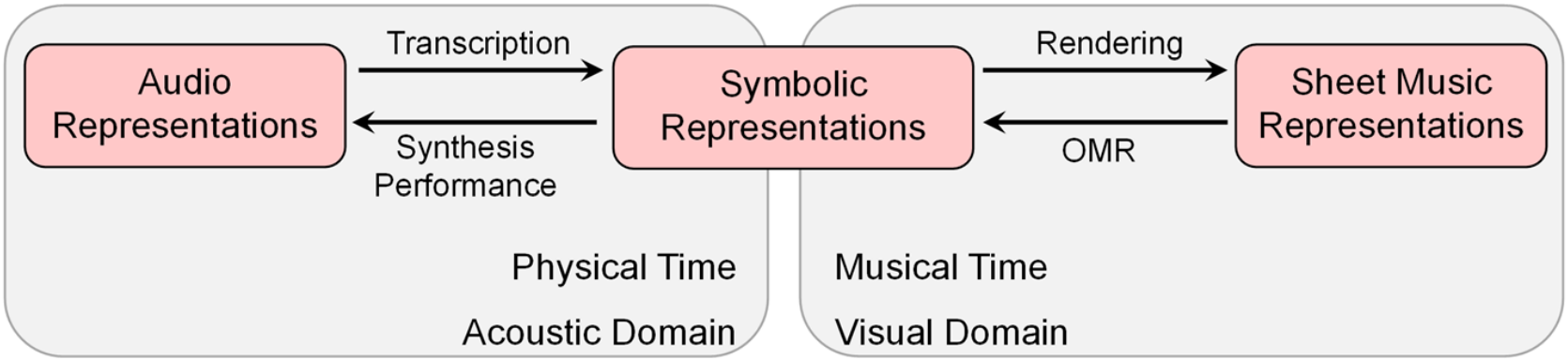
2023

Queen Mary University of London

Week 1 Recap

Music Representations

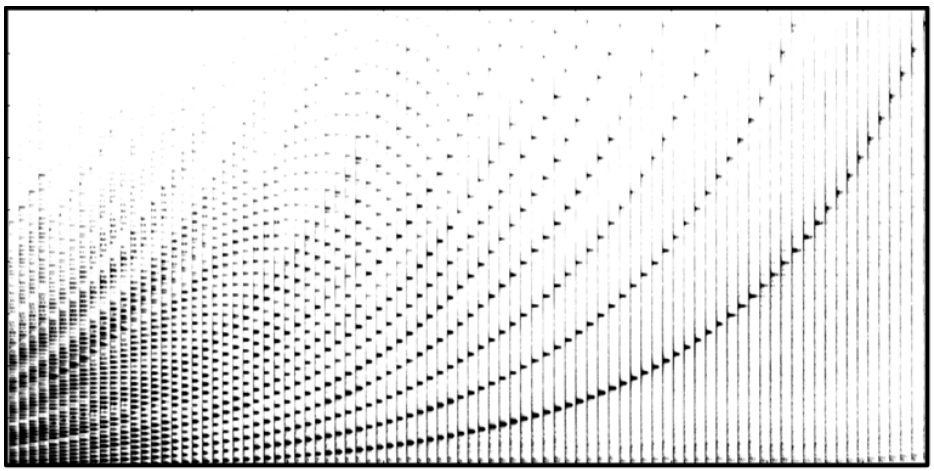
- Sheet music representations
- Symbolic representations
- Audio representations



This week's content

Time-Frequency Representations

- Signals
- Fourier Transform
- Discrete Fourier Transform (DFT)
- Short-Time Fourier Transform (STFT)
- Perceptually-motivated representations



Signals

Signals

Signal

A signal is a function that conveys information about the state or behaviour of a physical system.

- Example: time-varying sound pressure, motion of a particle, distribution of light, sequence of images...
- Here, we consider audio signals, which depict the amplitude of air pressure over time.
- Two different types of signals: analog and digital

Analog Signals

Analog Signal

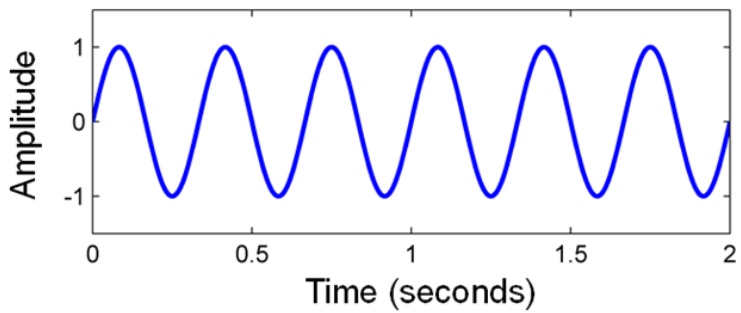
An analog signal is a function $f : \mathbb{R} \to \mathbb{R}$ which assigns an amplitude value $f(t) \in \mathbb{R}$ to each time point $t \in \mathbb{R}$.

Periodic Signal

A signal f is called periodic with period $\lambda \in \mathbb{R}_{>0}$ if $f(t) = f(t + \lambda)$ holds for all $t \in \mathbb{R}$.

Sinusoid

A periodic function defined by $f(t) := A \sin(2\pi(\omega t - \phi))$.
A: amplitude; ω: frequency; ϕ: phase

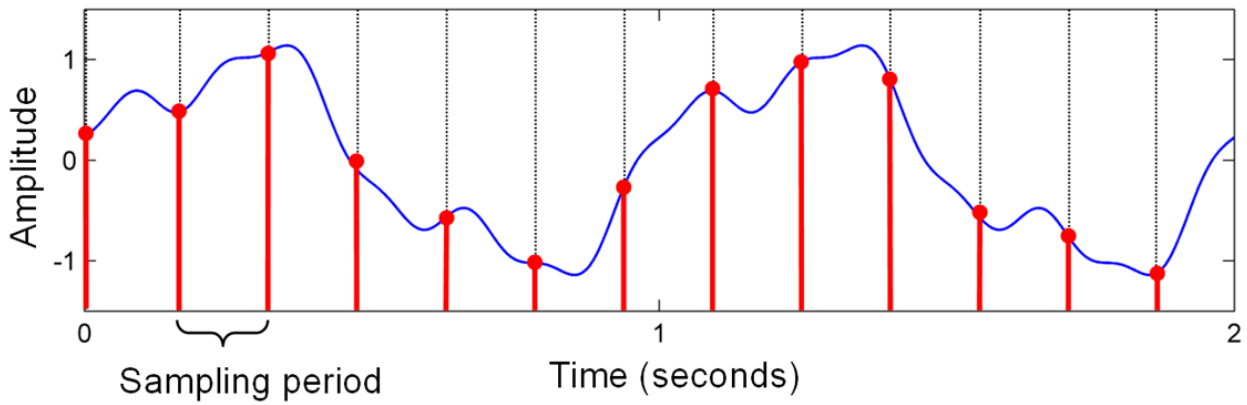


Digital Signals

- Computers can only store and process a finite number of values, therefore we need to convert waveforms into a discrete representation ("digitisation").
- Digitisation typically involves two steps: sampling and quantisation.

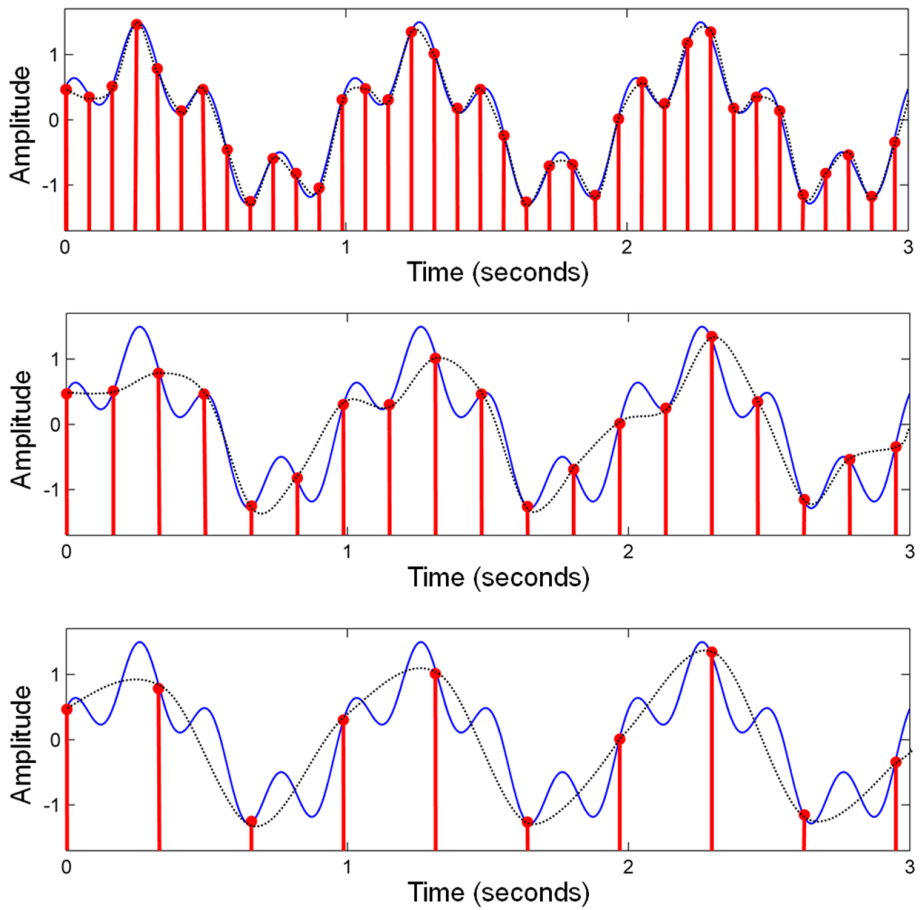
Sampling

The process of reducing a continuous-time signal to a discrete-time signal: $x(n) := f(n \cdot T)$, where T is the sampling period and $n \in \mathbb{Z}$.



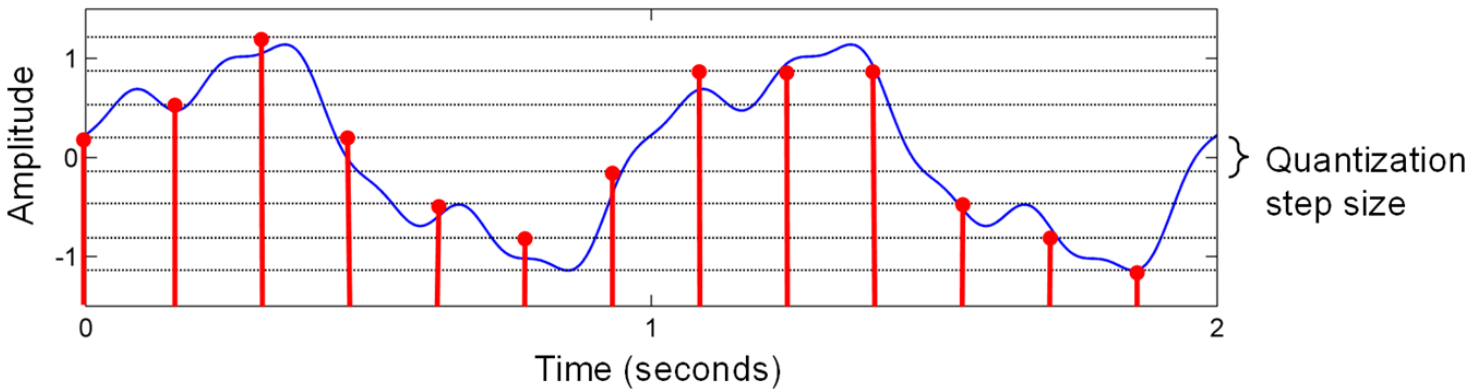
Digital Signals

- Sampling is a lossy operation, in that the original analog signal cannot be recovered from its sampled version.
- Sampling might cause an effect known as aliasing, where certain frequency components of the signal become indistinguishable.



Quantisation

Replacing the continuous range of possible amplitudes by a discrete range of possible values (typically rounding off to some unit of precision).

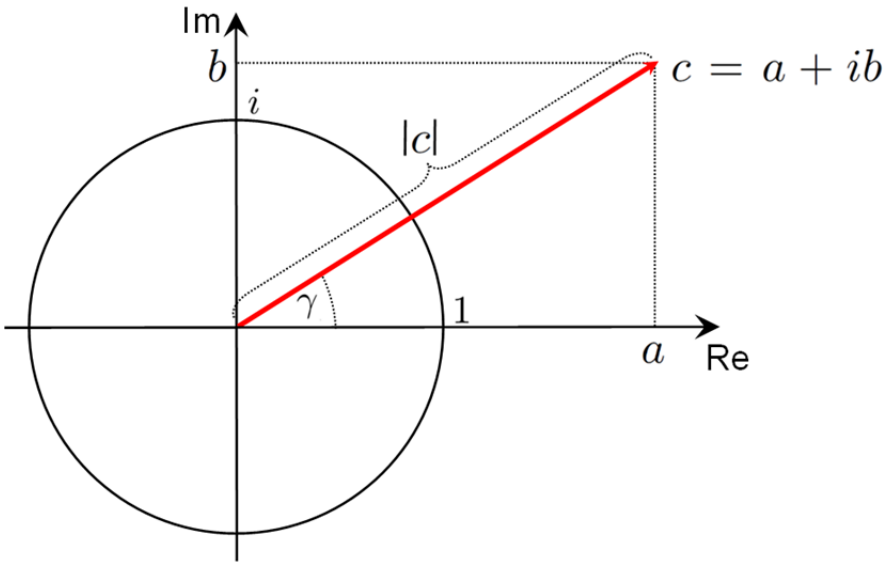


- The difference between the actual analog value and the quantised value is called the quantisation error.

Intermission: Complex numbers

- Complex numbers extend real numbers by introducing the imaginary number $i := \sqrt{-1}$ with the property $i^2 = -1$.
- Each complex number can be written as $c = a + ib$, where $a \in \mathbb{R}$ is the real part and $b \in \mathbb{R}$ the imaginary part of c.
- The set of all complex numbers is written as $\mathbb{C}$, and can be represented using polar coordinates:

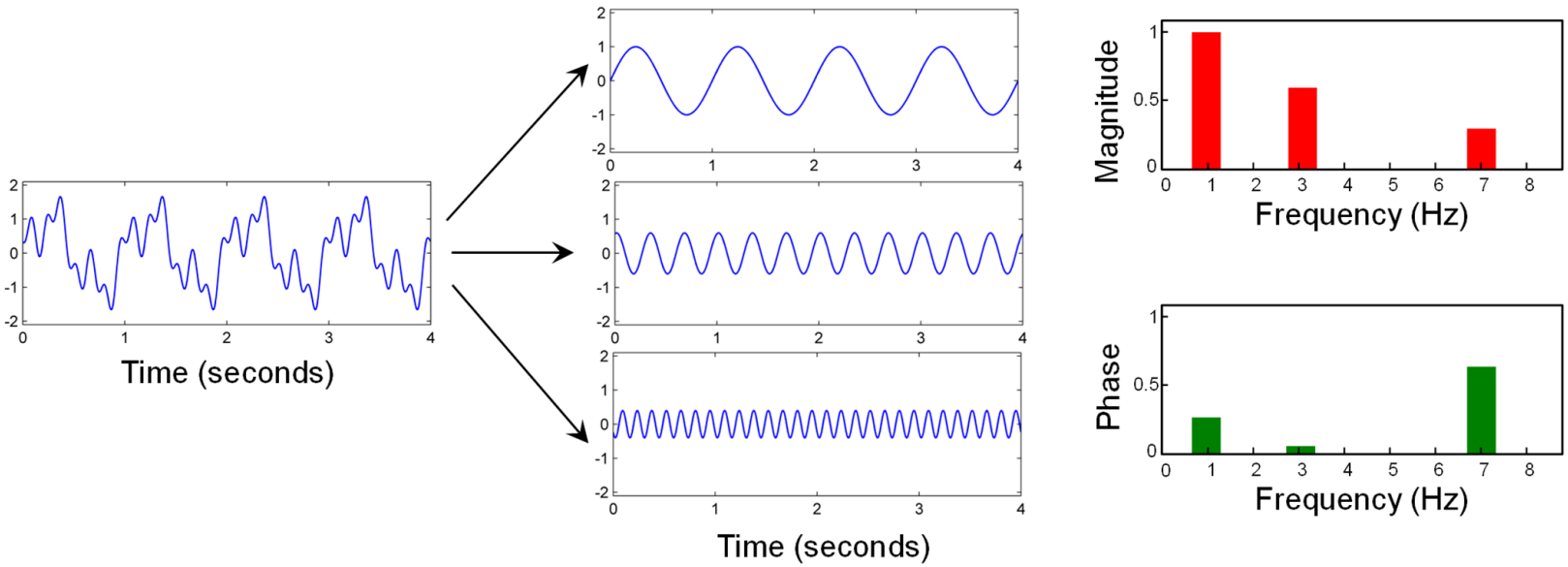
$$|c| := \sqrt{a^2 + b^2} \qquad \gamma := atan2(b, a)$$



Fourier Transform

Fourier Transform

Idea: Decompose a given signal into a superposition of sinusoids (elementary signals).

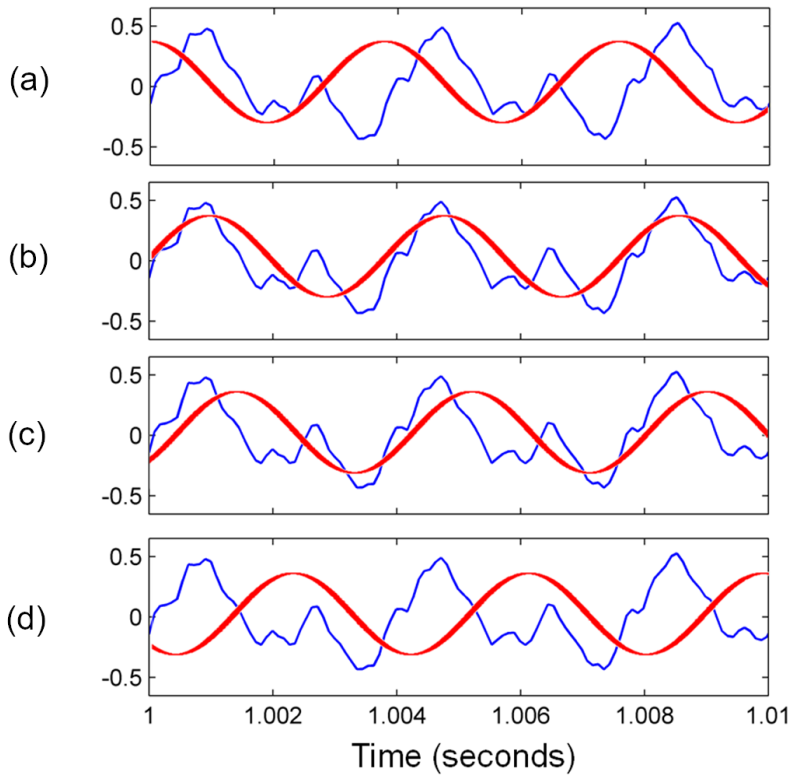


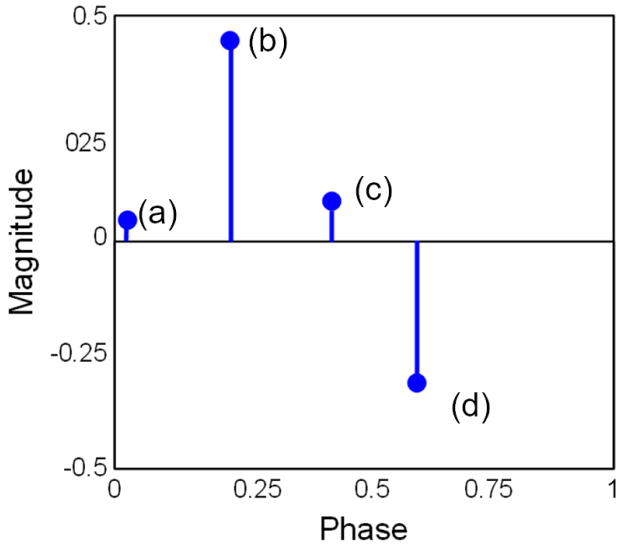
Interpretation:

- The magnitude A reflects the intensity at which the sinusoid of frequency ω appears in the signal.
- The phase ϕ reflects how the sinusoid has to be shifted to best correlate with the signal.

The Role of the Phase

The degree of similarity between the signal and a sinusoid of fixed frequency crucially depends on the phase.

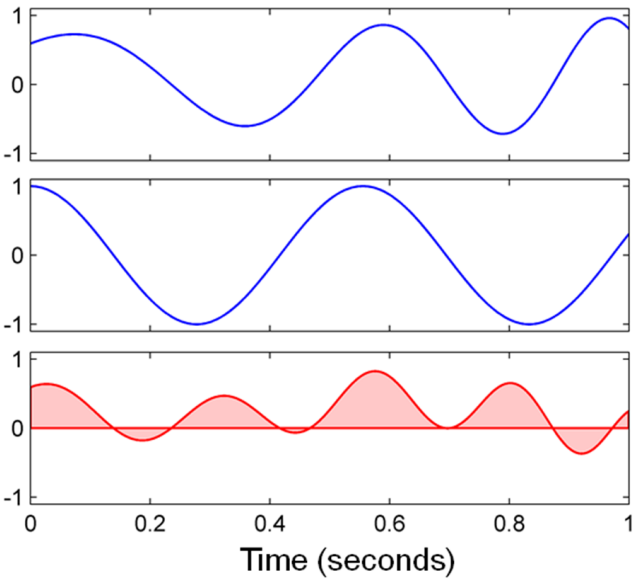


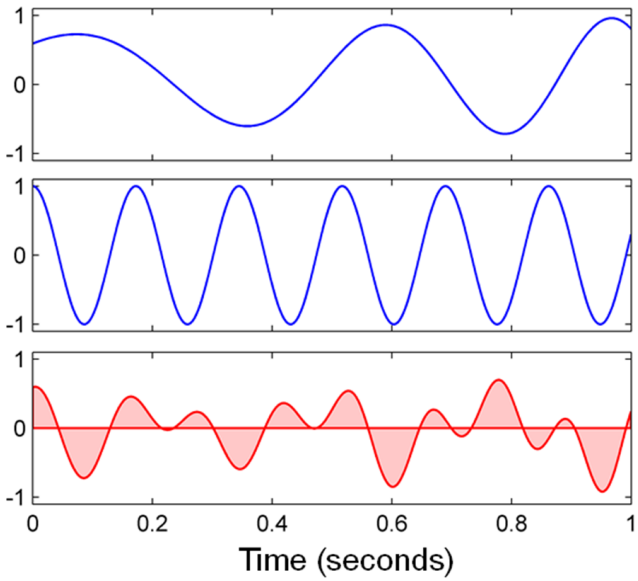


Computing similarity with integrals

- Assume two functions f and g - what does it mean for f and g to be similar?
- The joint behaviour of these functions can be captured by forming the integral of the product of the two functions:

$$\int_{t \in \mathbb{R}} f(t) \cdot g(t) dt$$





Fourier transform

$$c_\omega = \hat{f}(\omega) = \int_{t \in \mathbb{R}} f(t) \exp(-2\pi i \omega t) dt$$

The values c_ω are called the Fourier coefficients.

It can be seen that:

$$\hat{f}(\omega) = \int_{t \in \mathbb{R}} f(t) \cos(-2\pi\omega t) dt + i \int_{t \in \mathbb{R}} f(t) \sin(-2\pi\omega t) dt$$

The absolute value $|\hat{f}(\omega)|$ is also called the magnitude of the Fourier coefficient.

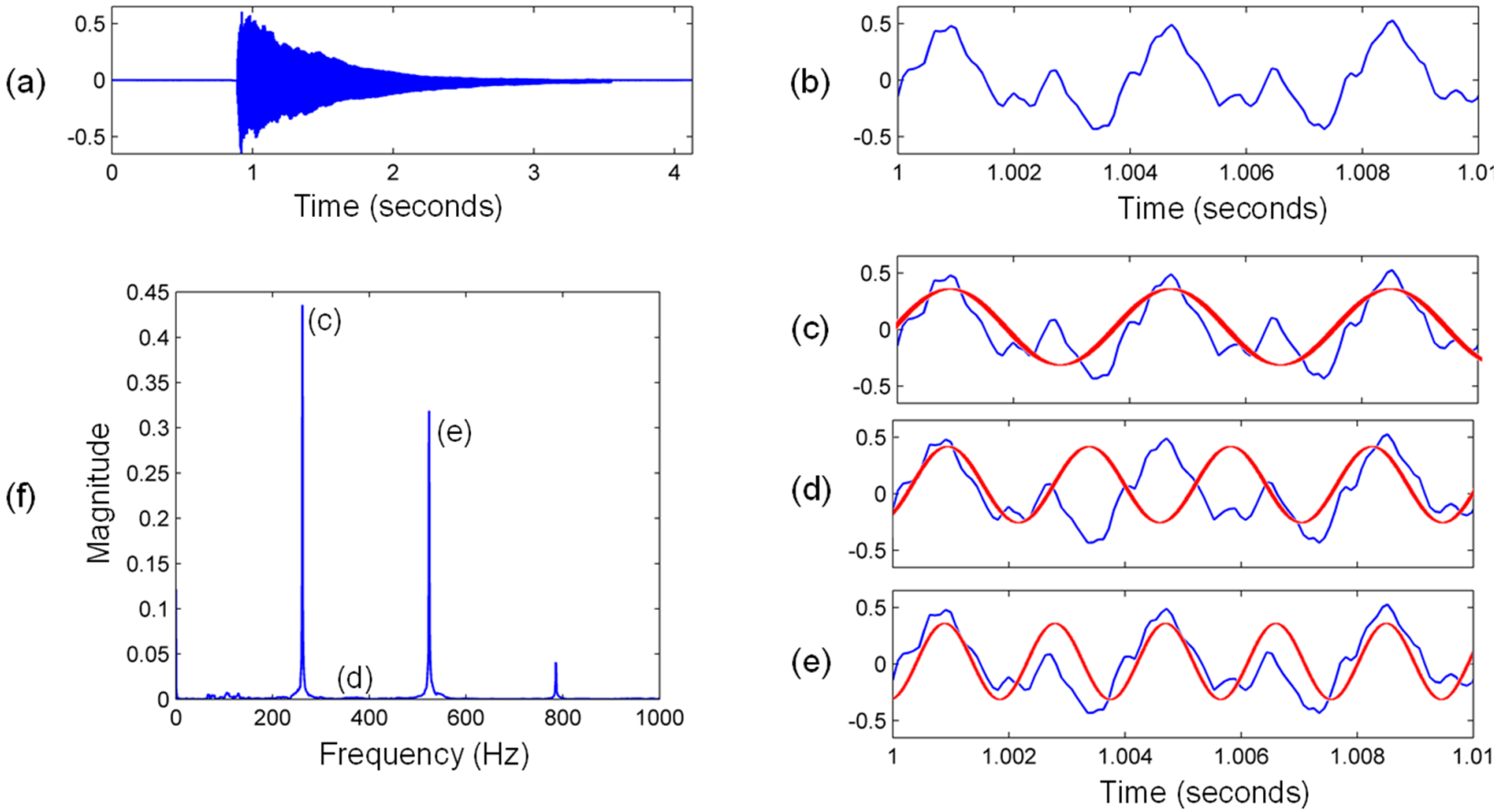
Fourier representation

The original signal can be reconstructed from its Fourier transform:

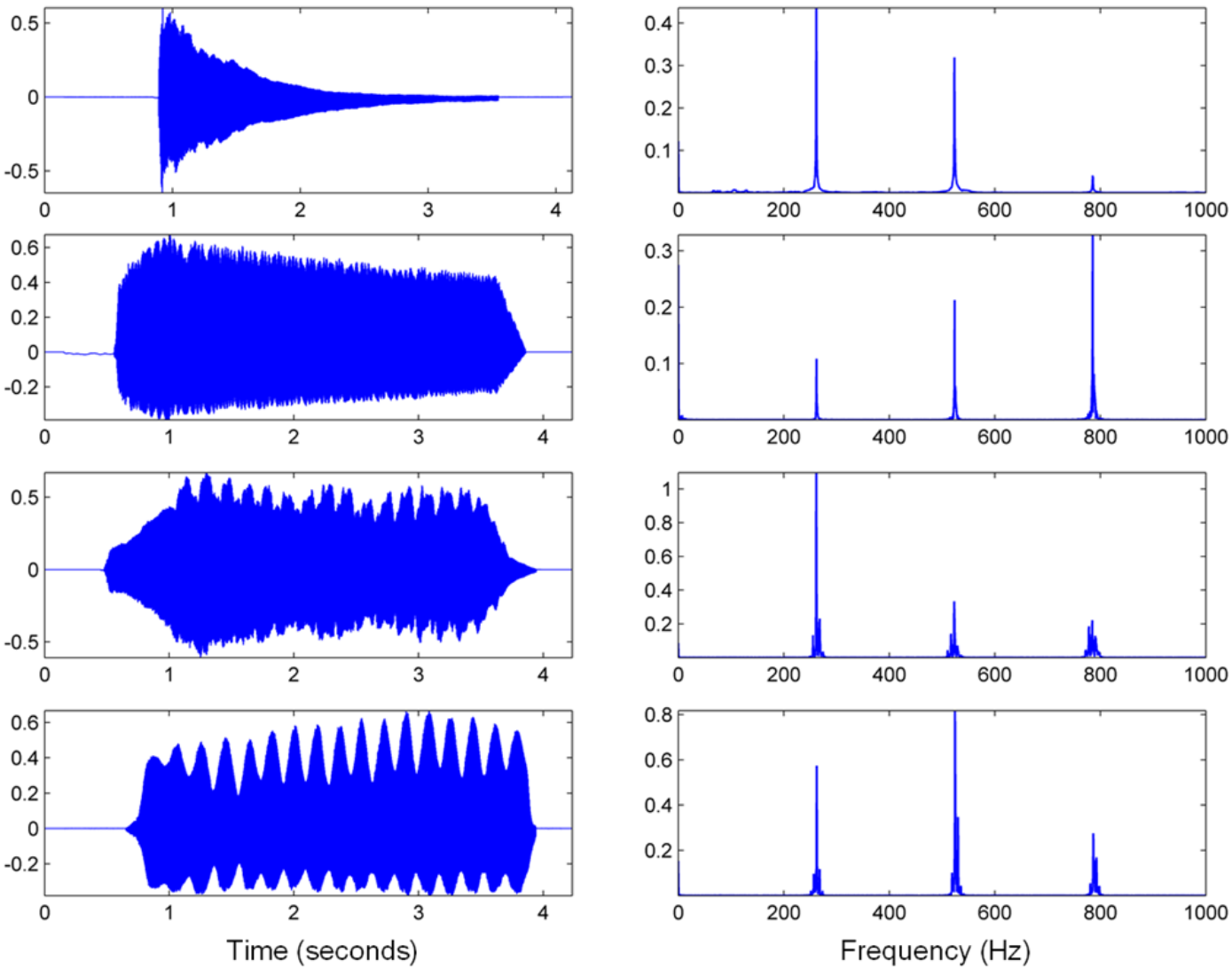
Fourier representation

$$f(t) = \int_{\omega \in \mathbb{R} \geqslant 0} c_\omega \exp(2\pi i \omega t) d\omega$$

Fourier transform examples



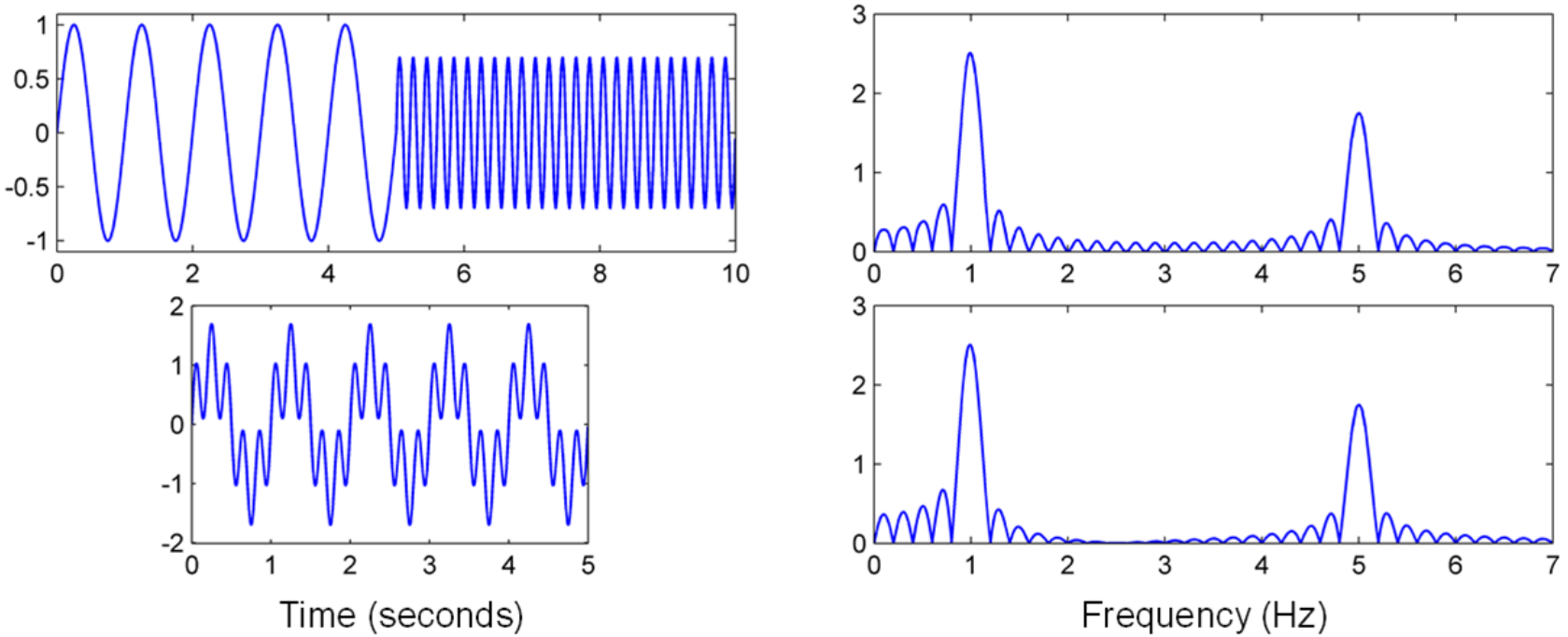
Fourier transform examples



Piano: Trumpet: Violin: Flute:

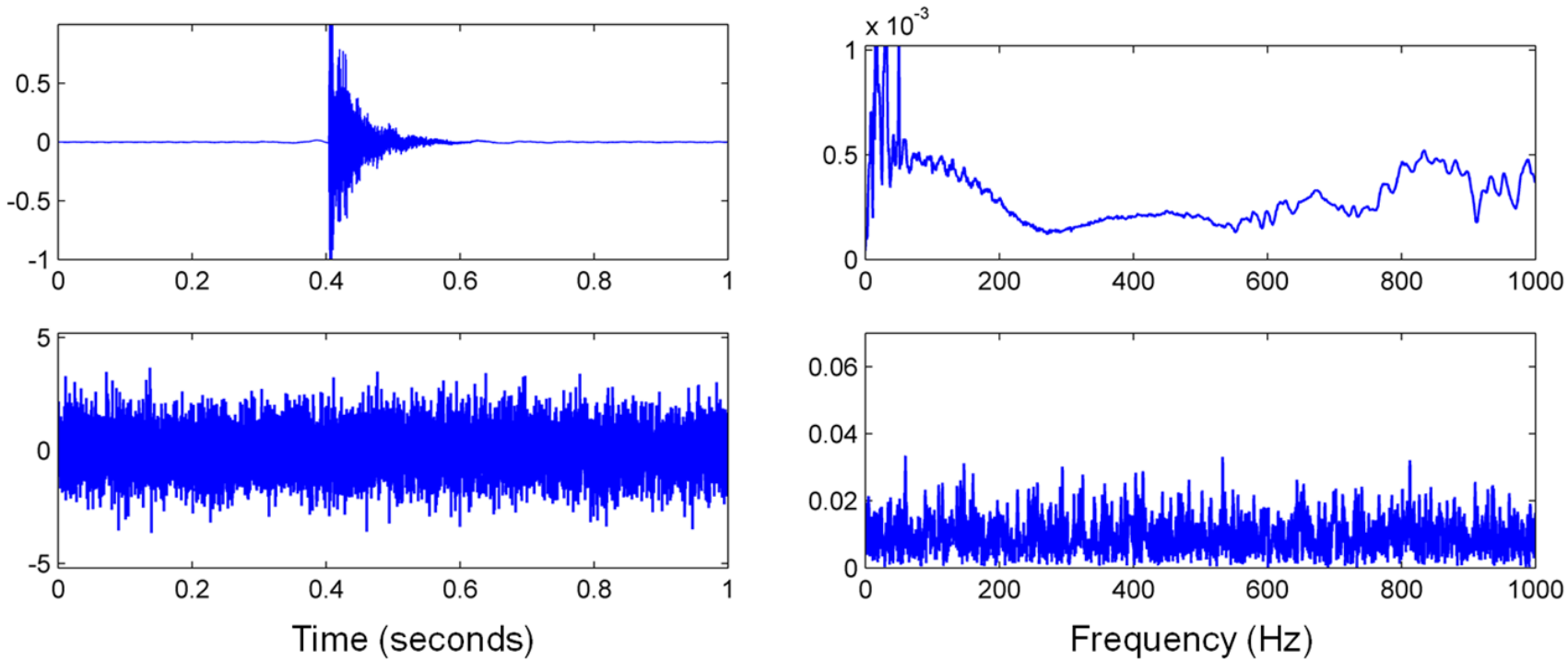
Fourier transform examples

The Fourier transform tells which frequencies occur, but does not tell when the frequencies occur.



Fourier transform examples

Waveform and Fourier transform of a clapping sound and white noise:



Discrete Fourier Transform

Discrete Fourier Transform

- Making the Fourier transform work with discrete signals $x(n) := f(n \cdot T)$, where $x(n)$: sample; $F_s := 1/T$: sampling rate

Sampling theorem

The original analog signal f can be reconstructed perfectly from its sampled version x if f does not contain any frequencies higher than $\Omega := F_s/2$ Hz.

(if it does, sampling would cause aliasing)

- In the following, we assume that f is bandlimited and that f has a finite duration.

Discrete Fourier Transform

Assume a discrete signal $x(n)$ with relevant samples $x(0), x(1), \ldots, x(N-1)$.

Discrete Fourier Transform (DFT)

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp(-2\pi i k n / N)$$

where $k \in [0 : M-1]$, $M \in \mathbb{N}$.

- Linking indices k of $X(k)$ with physical frequencies:

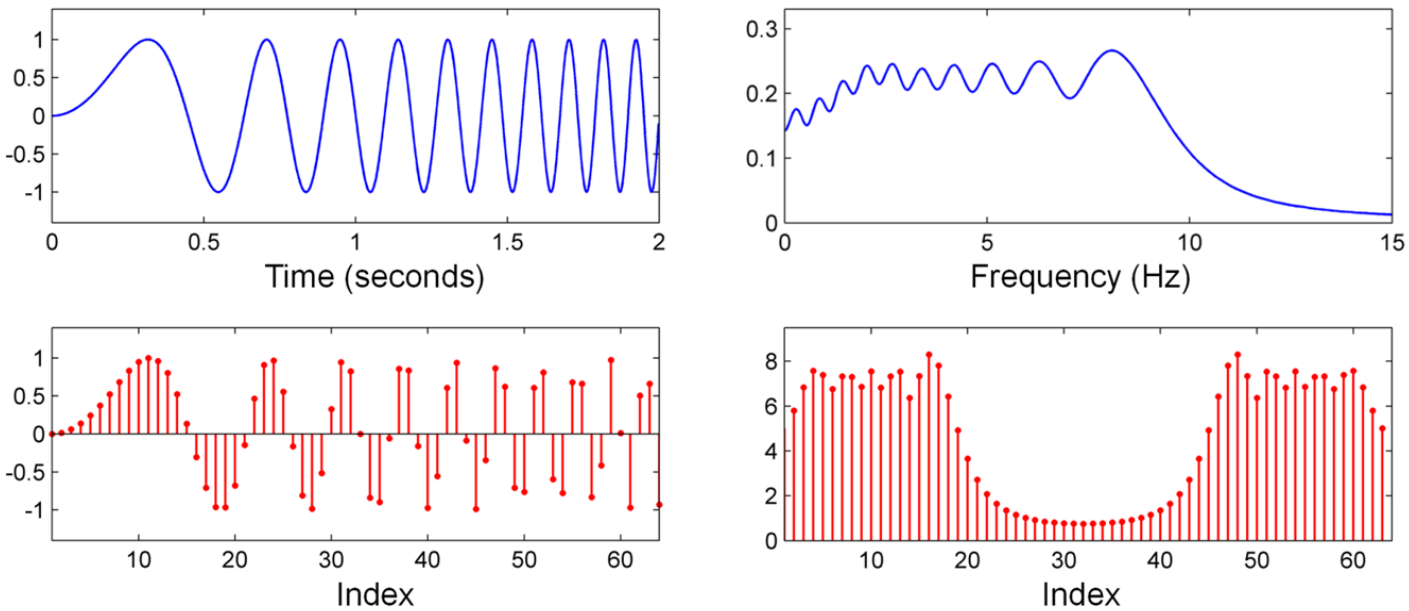
$$F_{\text{coef}}(k) := \frac{k \cdot F_s}{N} \text{ Hz}$$

Discrete Fourier Transform

Linking the DFT with the Fourier transform:

$$X(k) \approx \frac{1}{T}\hat{f}\left(\frac{k}{N} \cdot \frac{1}{T}\right)$$

Given that $X(k) = \overline{X(N-k)}$, coefficients $X(k)$ are redundant for $k = \left\lfloor \frac{N}{2} \right\rfloor + 1, \ldots, N-1$



Fast Fourier Transform

The DFT requires $O(N^2)$ multiplications and additions.
Solution: Fast Fourier Transform (FFT) - a recursive algorithm with $O(N \log N)$ multiplications and additions.

Algorithm: FFT

Input: The length $N = 2^L$ with N being a power of two
The vector $(x(0),\ldots,x(N-1))^\top \in \mathbb{C}^N$

Output: The vector $(X(0),\ldots,X(N-1))^\top = \mathrm{DFT}_N \cdot (x(0),\ldots,x(N-1))^\top$

Procedure: Let $(X(0),\ldots,X(N-1)) = \mathrm{FFT}(N,x(0),\ldots,x(N-1))$ denote the general form of the FFT algorithm.
If $N = 1$ then

$$X(0) = x(0).$$

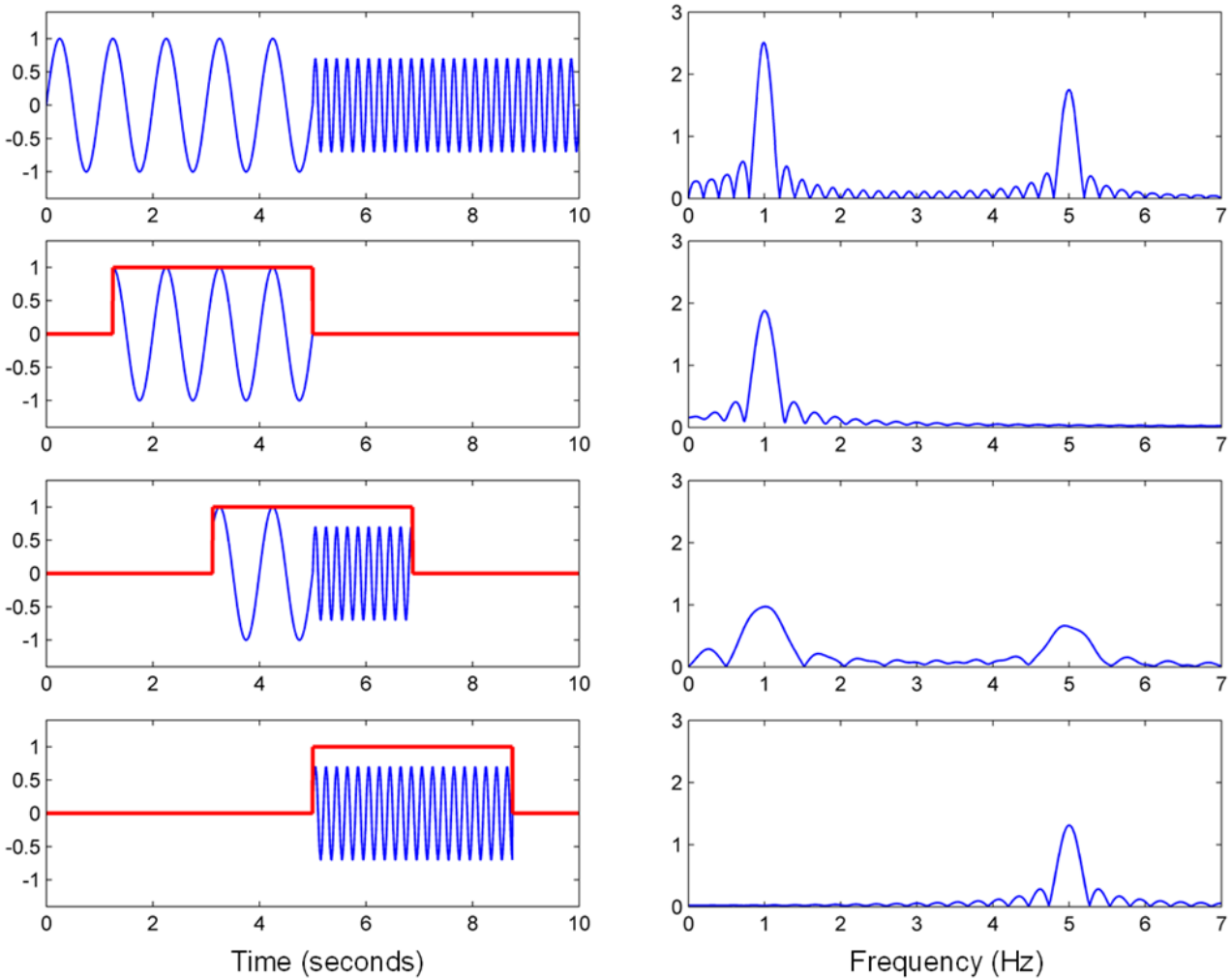
Otherwise compute recursively:

$$(A(0),\ldots,A(N/2-1)) = \mathrm{FFT}(N/2,x(0),x(2),x(4)\ldots,x(N-2)),$$
$$(B(0),\ldots,B(N/2-1)) = \mathrm{FFT}(N/2,x(1),x(3),x(5),\ldots,x(N-1)),$$
$$C(k) = \omega_N^k \cdot B(k) \text{ for } k \in [0:N/2-1],$$
$$X(k) = A(k) + C(k) \text{ for } k \in [0:N/2-1],$$
$$X(N/2+k) = A(k) - C(k) \text{ for } k \in [0:N/2-1].$$

Short-Time Fourier Transform

Short-Time Fourier Transform (STFT)

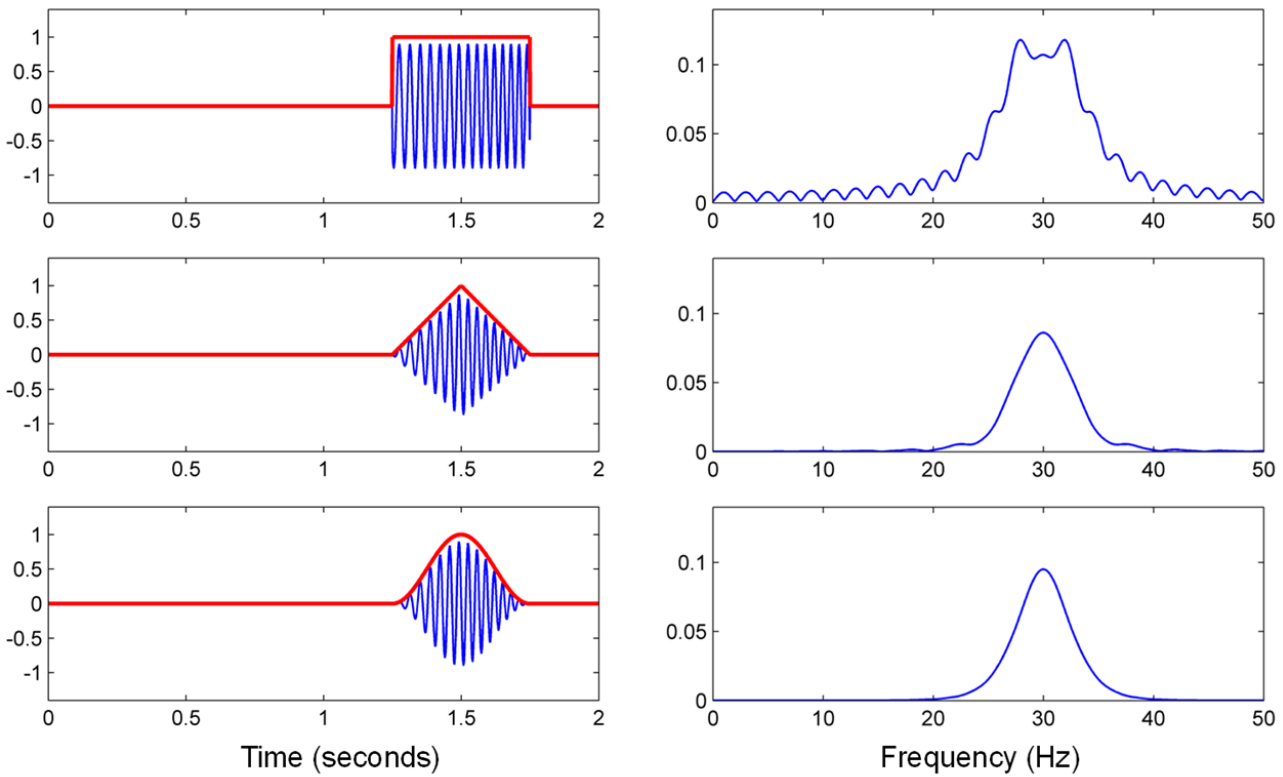
Idea: Consider only a small section of the signal for the spectral analysis
$\rightarrow$ recovery of time information



Short-Time Fourier Transform (STFT)

Formally, we consider a window function; and multiply the original signal with the window function to yield a windowed signal.

However: the STFT reflects not only the properties of the original signal but also those of the window function (length and shape of the window).



Short-Time Fourier Transform (STFT)

As with the Fourier transform, there is an analog and a discrete version of the STFT.

Discrete STFT

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

where w is a window function of length N, $m \in \mathbb{Z}$ denotes the time frame, and $H \in \mathbb{N}$ is the hop size.

So, $\mathcal{X}(m,k)$ denotes the kth Fourier coefficient for the mth time frame.

For each time frame m, one obtains a spectral vector which can be computed with the FFT.

Short-Time Fourier Transform (STFT)

- Linking STFT coefficients with time positions:

$$T_{\text{coef}}(m) := \frac{m \cdot H}{F_s} \text{ sec}$$

- Linking STFT coefficients with physical frequencies:

$$F_{\text{coef}}(k) := \frac{k \cdot F_s}{N} \text{ Hz}$$

- A common choice for hop size is $H = N/2$, as a trade-off between temporal resolution and data volume.

Spectrogram

A two-dimensional representation of the squared magnitude of the STFT:

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

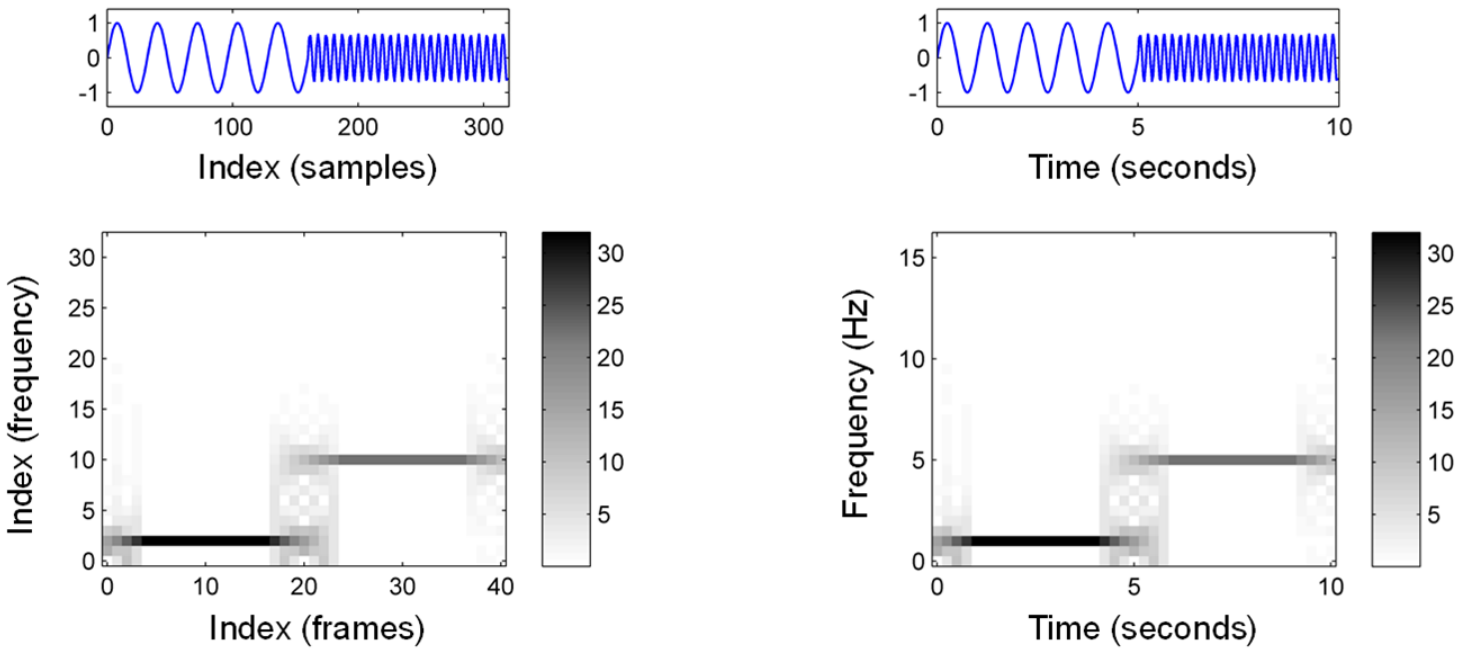
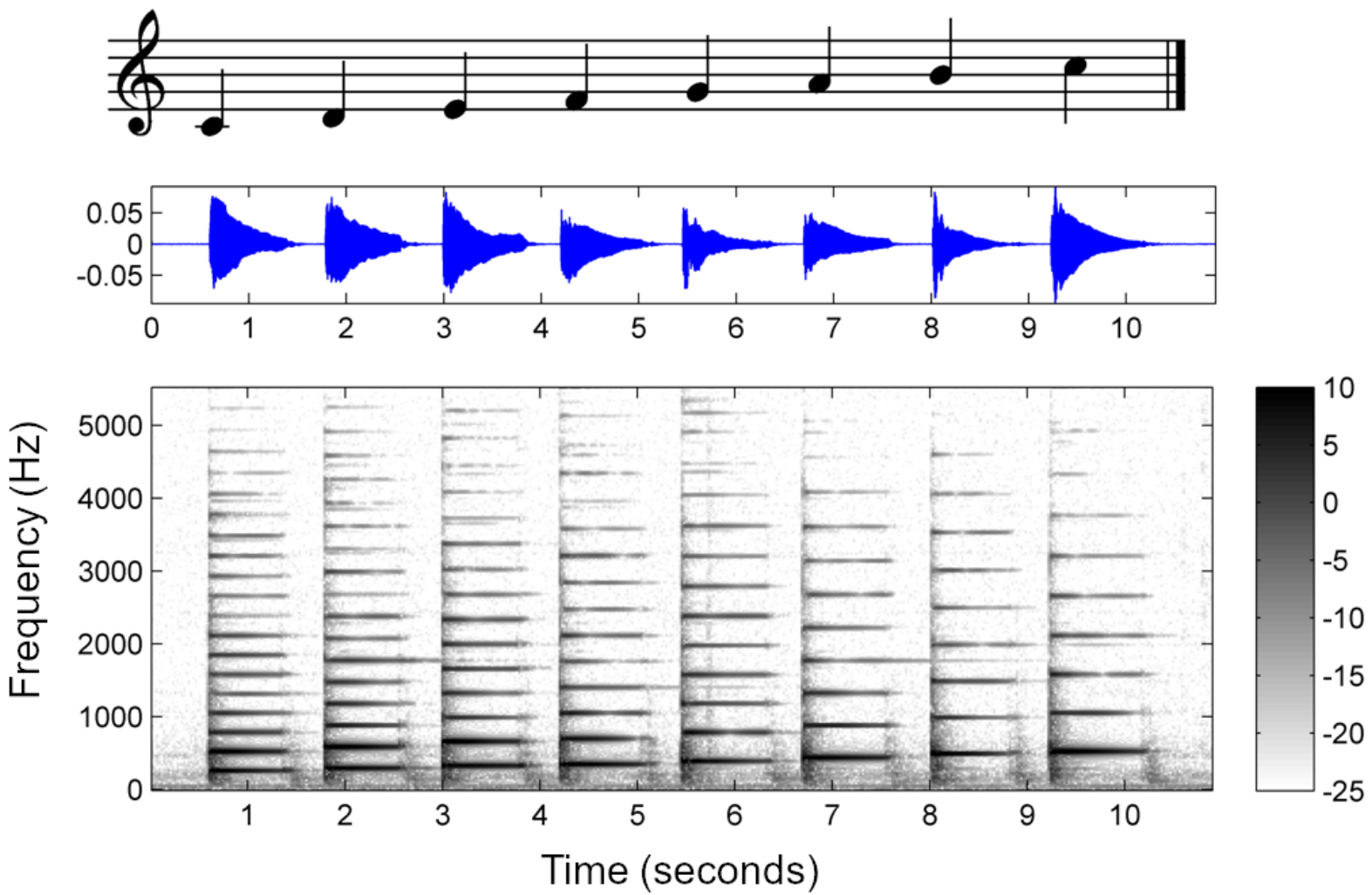


Figure: signal and corresponding spectrogram expressed in terms of samples/frames and seconds, respectively.

Spectrogram



Spectrogram

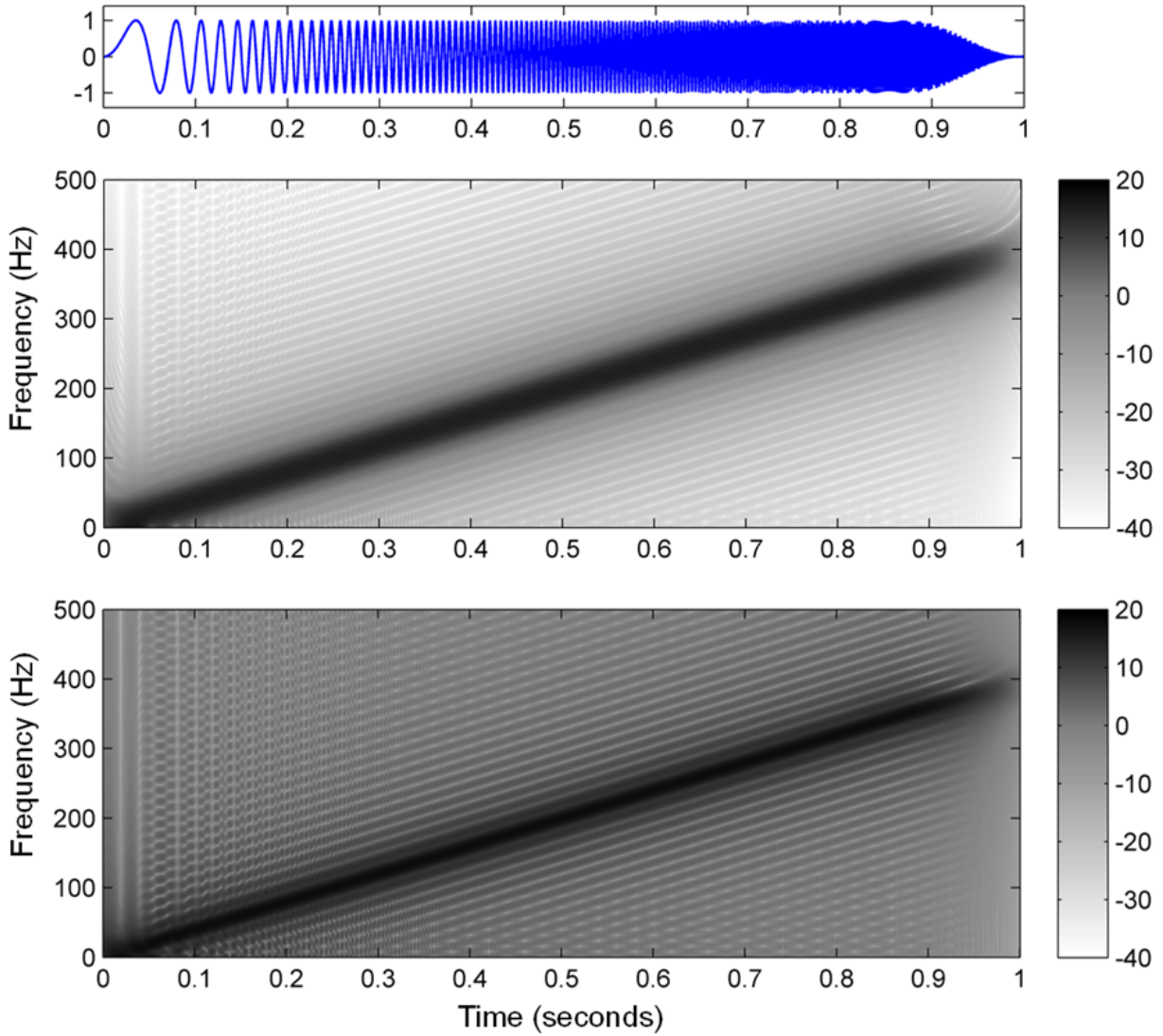


Figure: Effect of window type (Hann and rectangular).

Spectrogram

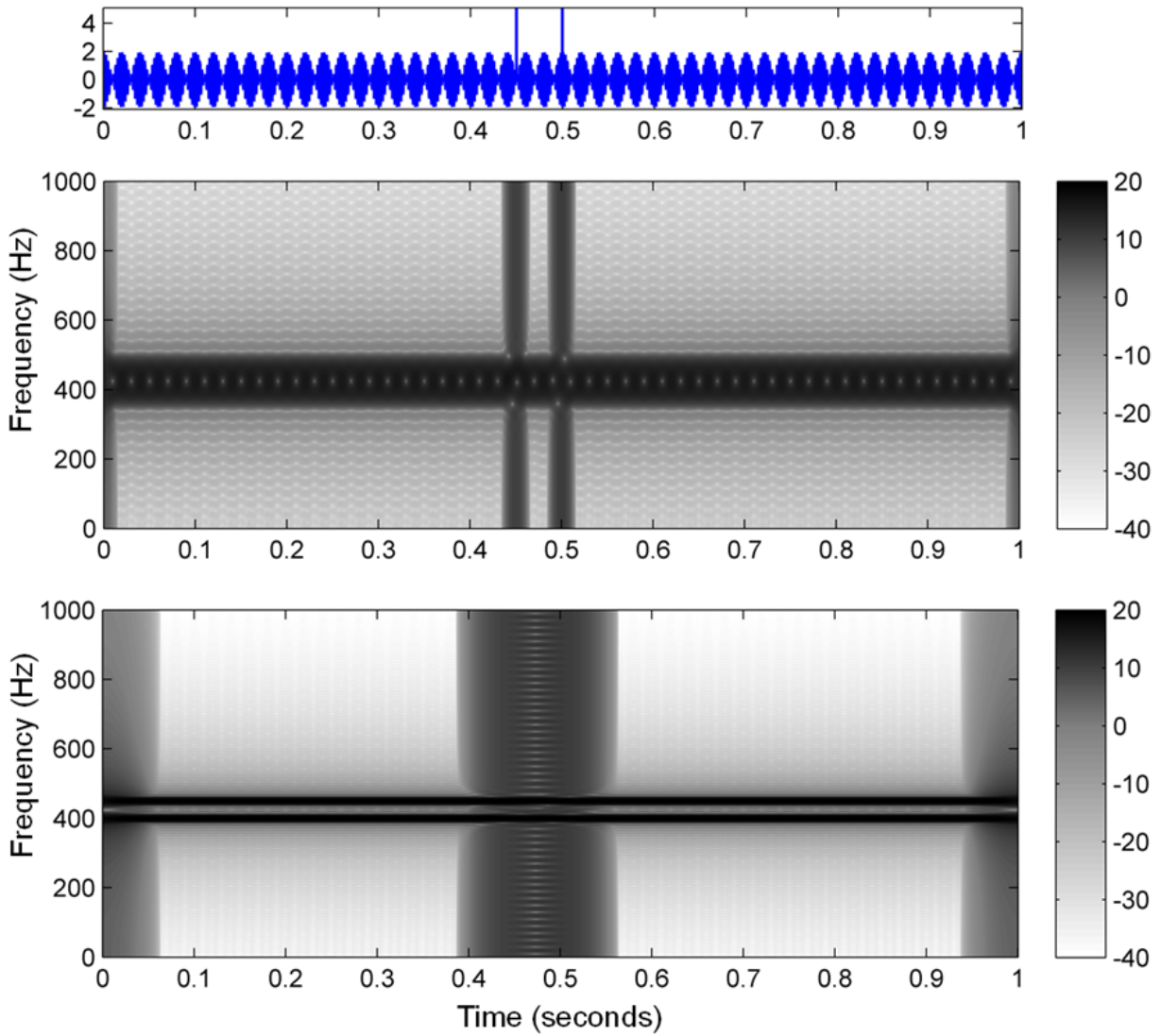


Figure: Effect of window size (32ms and 128ms).

Spectrogram

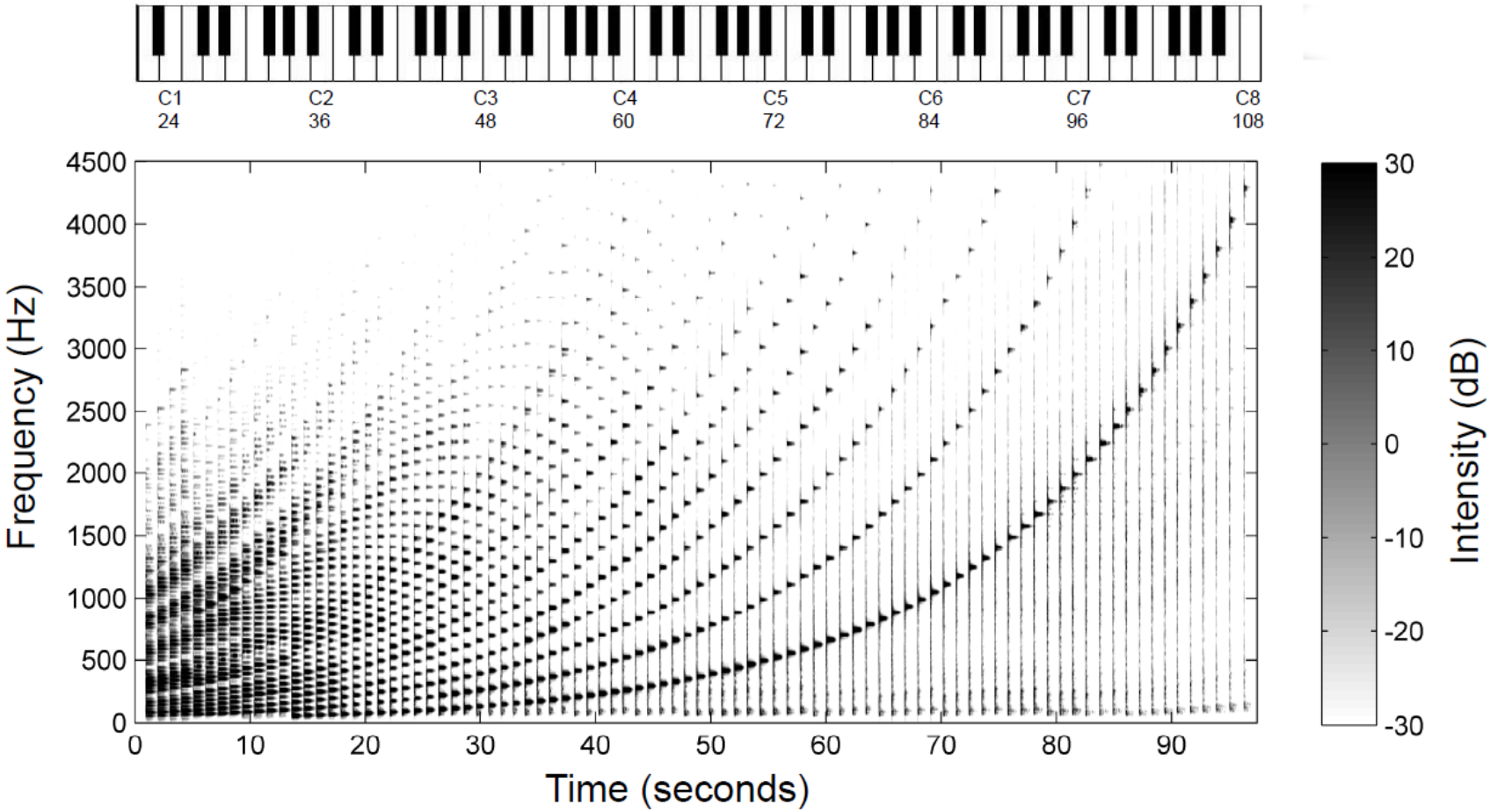


Figure: Chromatic scale.

Spectrogram

Time-Frequency Localisation

Size of window constitutes a trade-off between time localisation and frequency localisation:

- Large window: poor time localisation, good frequency localisation
- Small window: good time localisation, poor frequency localisation

A variant of the Heisenberg Uncertainty Principle states that there is no window function that simultaneously localizes in time and frequency with arbitrary precision.

Perceptually-motivated representations

Log-frequency Spectrogram

Notes of the equal-tempered scale depend on their center frequencies in a logarithmic fashion:

$$F_{pitch}(p) = 2^{(p-69)/12} \cdot 440$$

This logarithmic perception of frequency motivates the use of a representation with a logarithmic frequency axis:

Log-frequency spectrogram

$$\mathcal{Y}_{LF}(n,p) := \sum_{k \in P(p)} |\mathcal{X}(n,k)|^2$$

where $P(p) := \{k : F_{pitch}(p-0.5) \leq F_{coef}(k) < F_{pitch}(p+0.5)\}$ and $\mathcal{Y}_{LF} : \mathbb{Z} \times [0:127]$.

Log-frequency Spectrogram

Note	p	$F_{\text{pitch}}(p)$	$F_{\text{pitch}}(p-0.5)$	$F_{\text{pitch}}(p+0.5)$	$\text{BW}(p)$
C4	60	261.63	254.18	269.29	15.11
C♯4	61	277.18	269.29	285.30	16.01
D4	62	293.66	285.30	302.27	16.97
D♯4	63	311.13	302.27	320.24	17.97
E4	64	329.63	320.24	339.29	19.04
F4	65	349.23	339.29	359.46	20.18
F♯4	66	369.99	359.46	380.84	21.37
G4	67	392.00	380.84	403.48	22.65
G♯4	68	415.30	403.48	427.47	23.99
A4	69	440.00	427.47	452.89	25.41
A♯4	70	466.16	452.89	479.82	26.93
B4	71	493.88	479.82	508.36	28.53
C5	72	523.25	508.36	538.58	30.23

Table: Note numbers, center frequencies, cutoff frequencies, and bandwidth.

Log-frequency Spectrogram

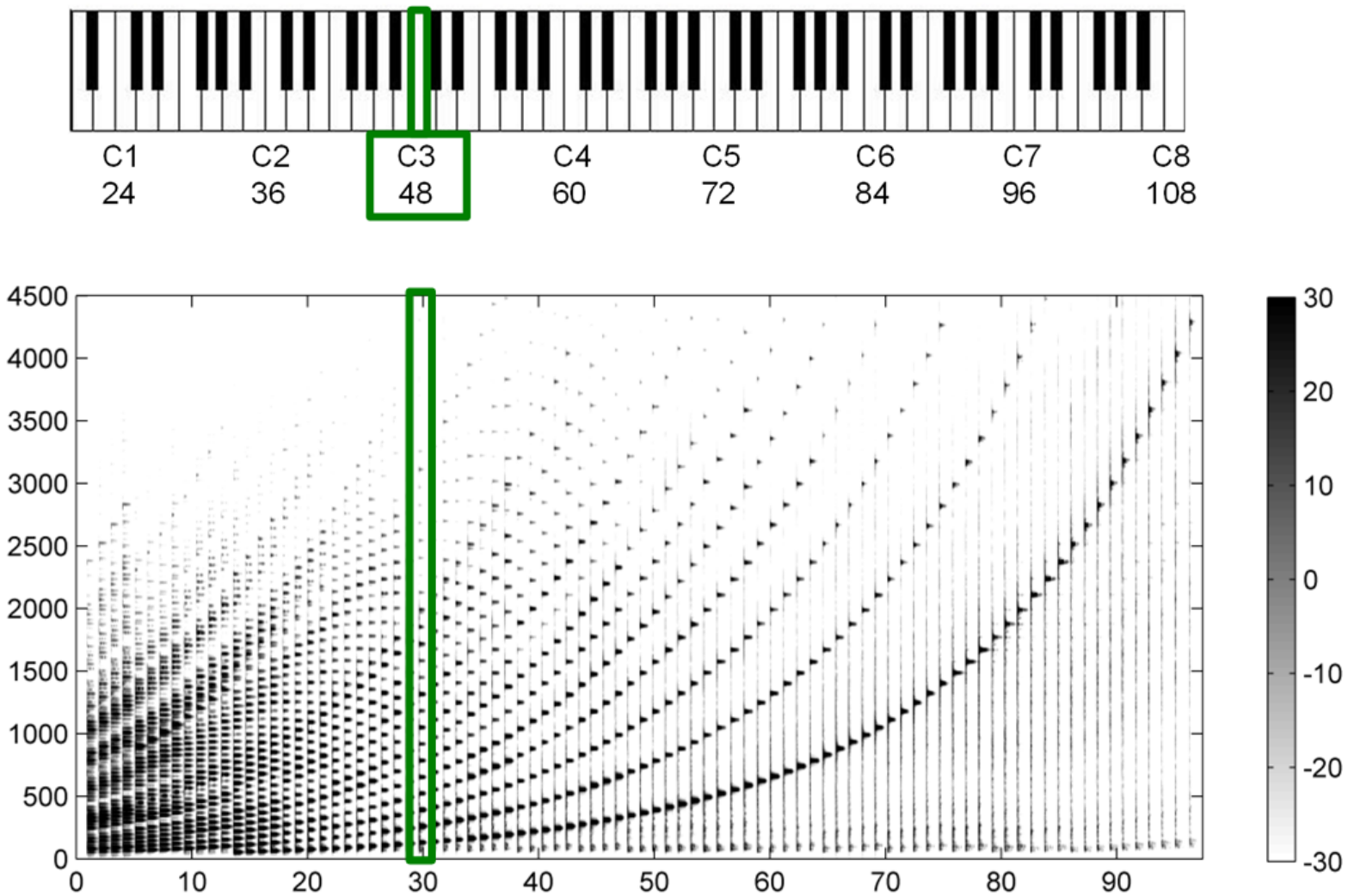


Figure: Linear frequency spectrogram for a chromatic scale.

Log-frequency Spectrogram

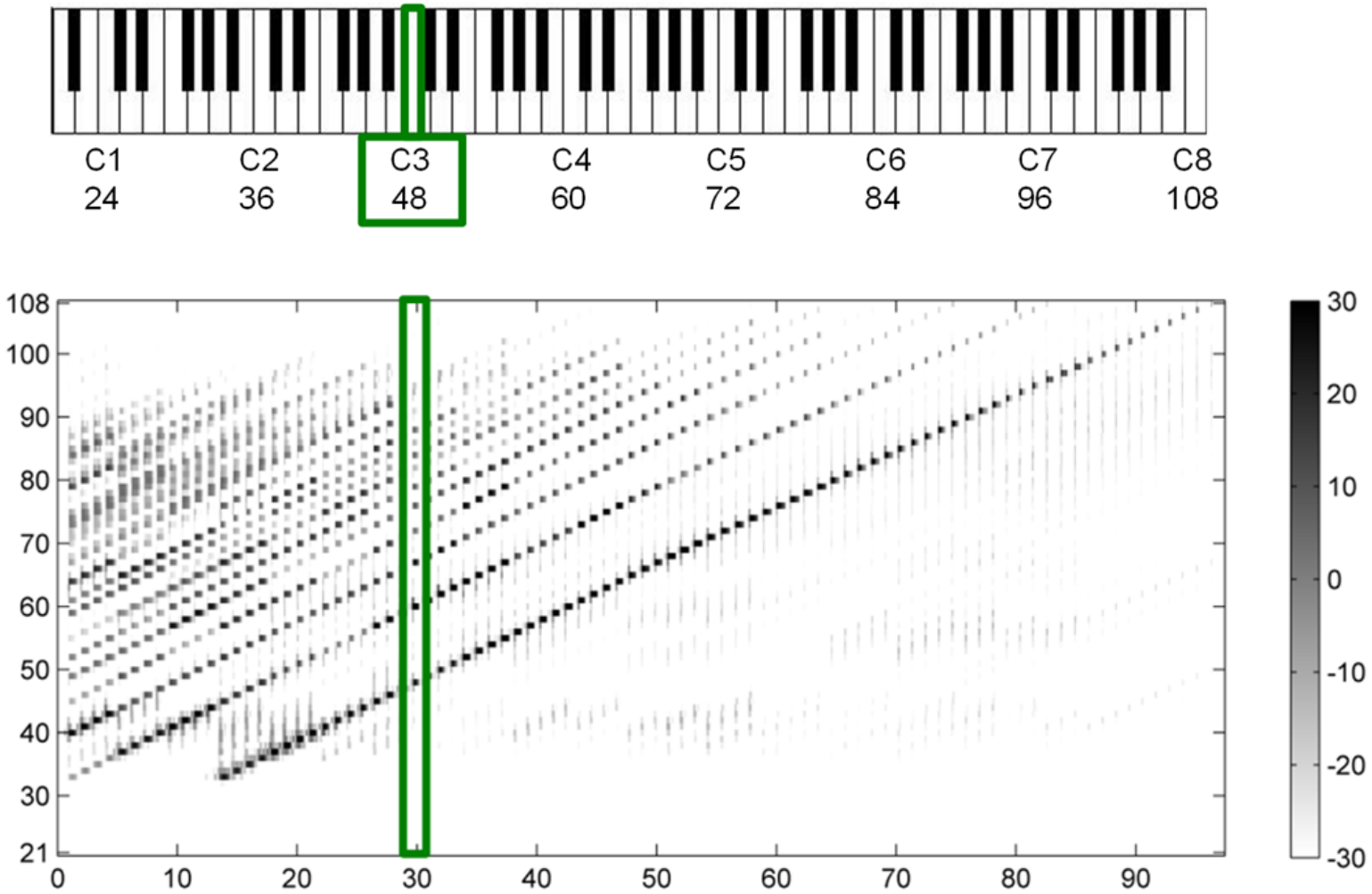


Figure: Log-frequency spectrogram for a chromatic scale.

Constant-Q Transform

STFT drawback: linearly spaced time/frequency bins
$\rightarrow$ constant time-frequency resolution

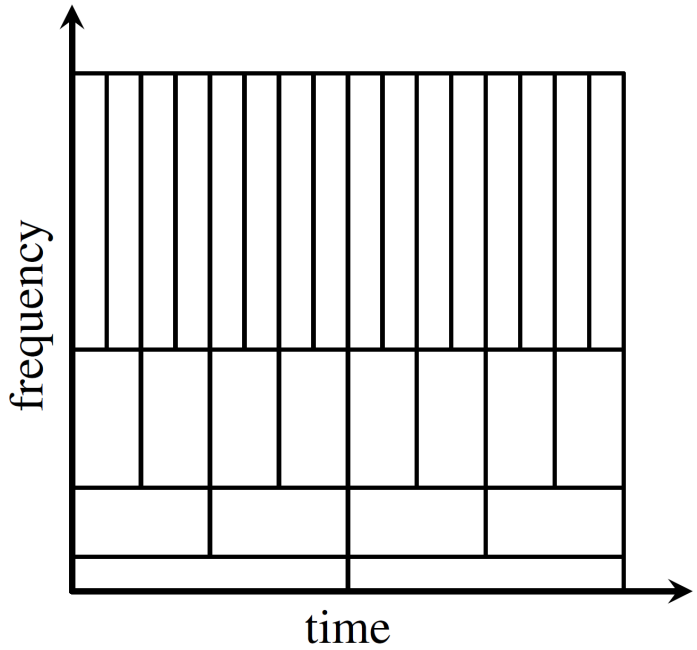
Constant-Q Transform (CQT): time-frequency representation where:

- Frequencies are logarithmically spaced
- Bins have a constant ratio of center frequencies over bandwidth (Q-factor)

In effect, that means that the frequency resolution is better for low frequencies and the time resolution is better for high frequencies

Also: musically and perceptually motivated representation

Constant-Q Transform



CQT drawbacks:

- CQT is computationally more intensive than the STFT/FFT
- CQT produces a data structure that is more difficult to handle than a linear frequency spectrogram

Constant-Q Transform

Log-frequency spectrogram vs. Constant-Q transform spectrogram

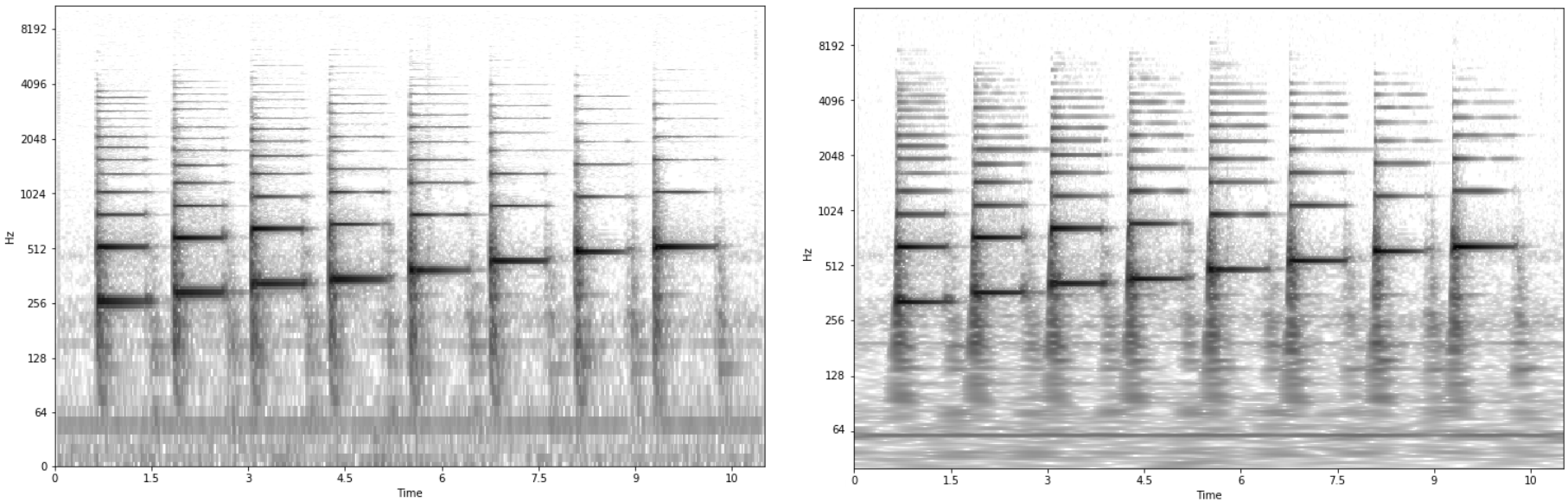


Figure: Left: log-frequency spectrogram; Right: CQT spectrogram.

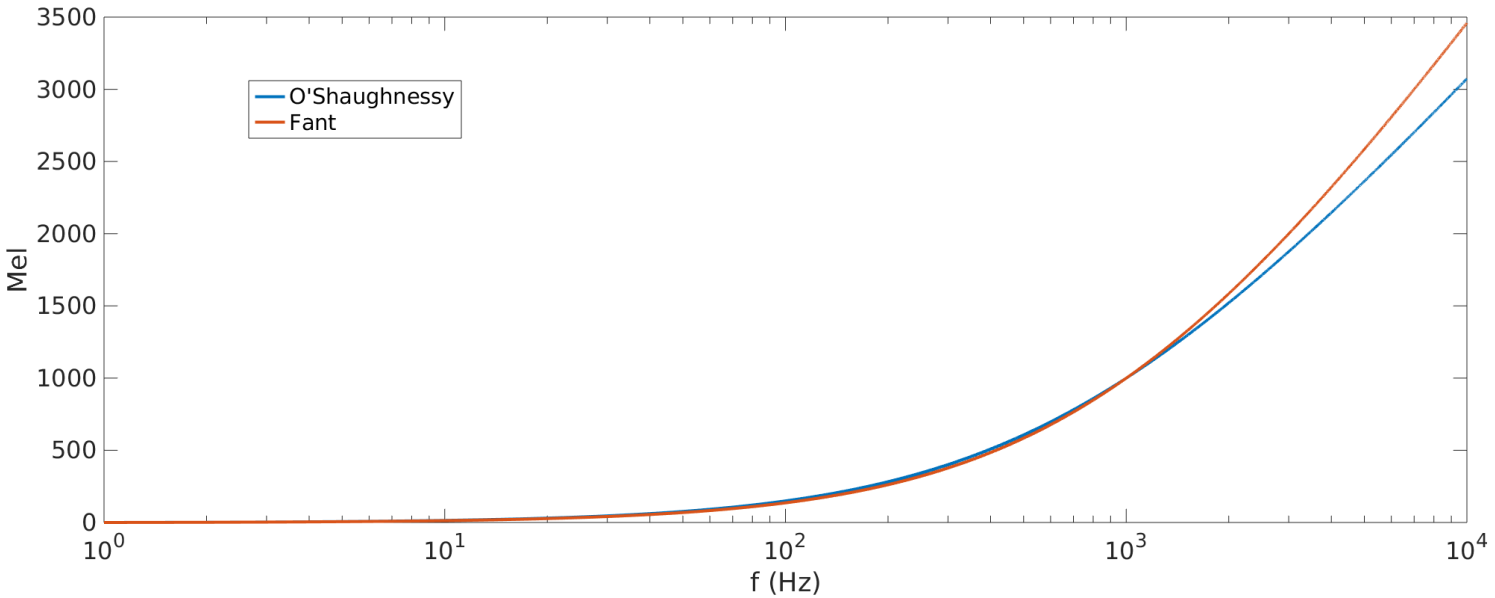
Mel scale

Mel scale

Perceptual scale of pitches judged by listeners to be equal in distance from one another:

$$m = 1000 \log_2 \left(1 + \frac{f}{1000}\right) \quad \text{or} \quad m = 2595 \log_{10} \left(1 + \frac{f}{700}\right)$$

as proposed by Fant (1973) and O'Shaughnessy (1987), respectively.



Mel spectrogram

Mel spectrogram: mapping an STFT spectrogram onto the mel scale.
$\rightarrow$ commonly used for deep learning-based MIR systems

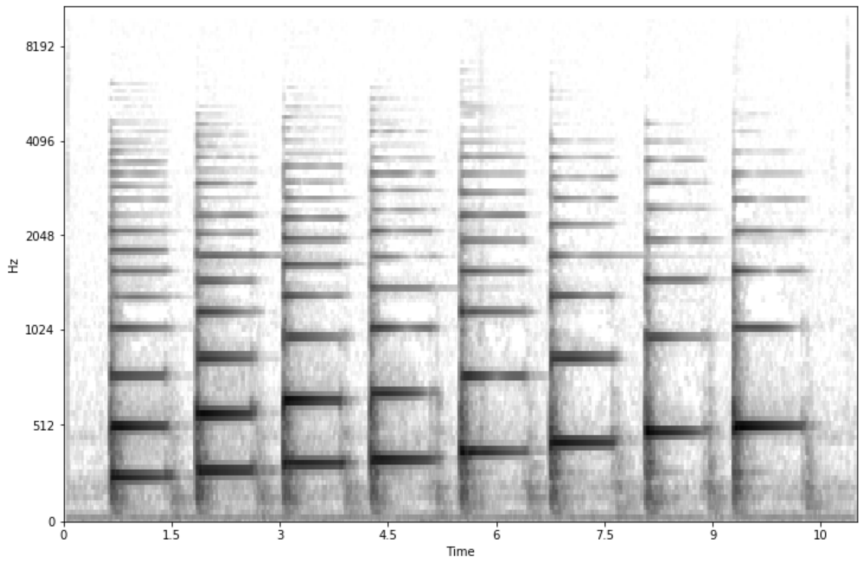


Figure: Mel spectrogram using 128 mel coefficients.

Other commonly used psychoacoustic pitch scales: Bark scale, Equivalent Rectangular Bandwidth (ERB) scale.

Time-frequency representations: Resources

- M. Müller, *Fundamentals of Music Processing*, Chapter 2 and section 3.1, Springer, 2015.
- J. C. Brown, *Calculation of a constant Q spectral transform*, Journal of the Acoustical Society of America, 89(1), 1991.
- C. Schörkhuber and A. Klapuri, *Constant-Q transform toolbox for music processing*, in Proc. Sound and Music Computing Conf., 2010.
- A. Lerch, *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, Section 5.1, Wiley, 2012.