Short-Time Fourier Transform Explained Easily

Valerio Velardo

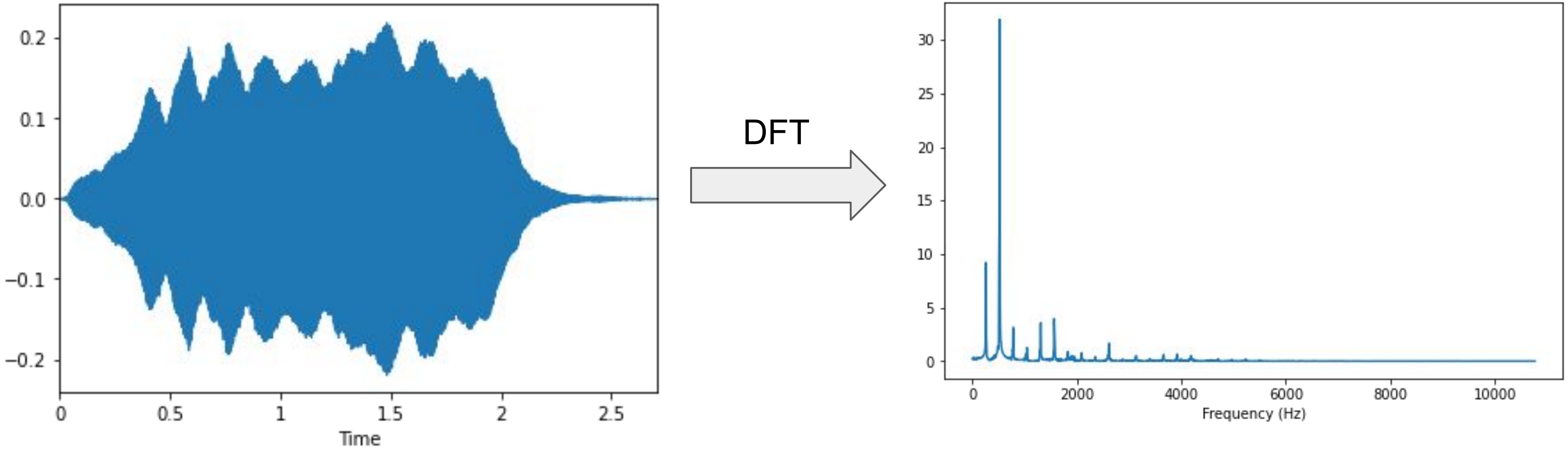
Join the community!

thesoundofai.slack.com

Previously...

$$\hat{x}(k/N) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

Previously...


0.2
0.1
0.0
−0.1
−0.2
0
0.5
1
1.5
2
2.5
Time
DFT
30
25
20
15
10
5
0
0
2000
4000
6000
8000
10000
Frequency (Hz)

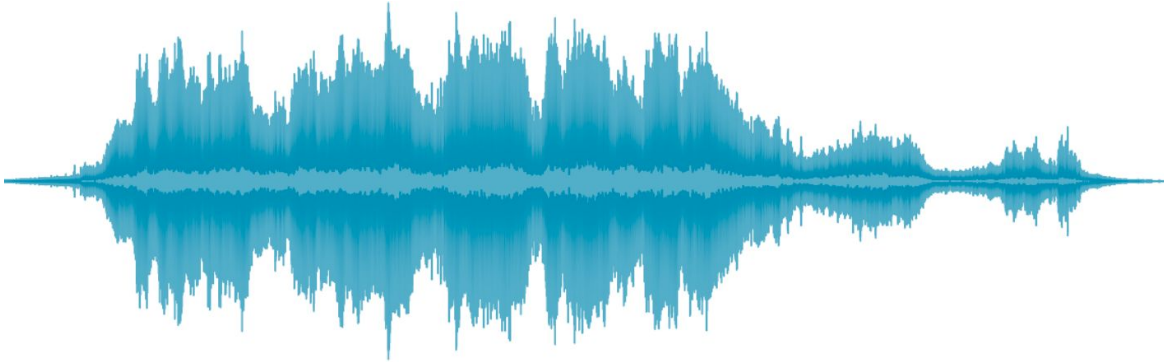
Fourier Transform Problem

WE KNOW WHAT

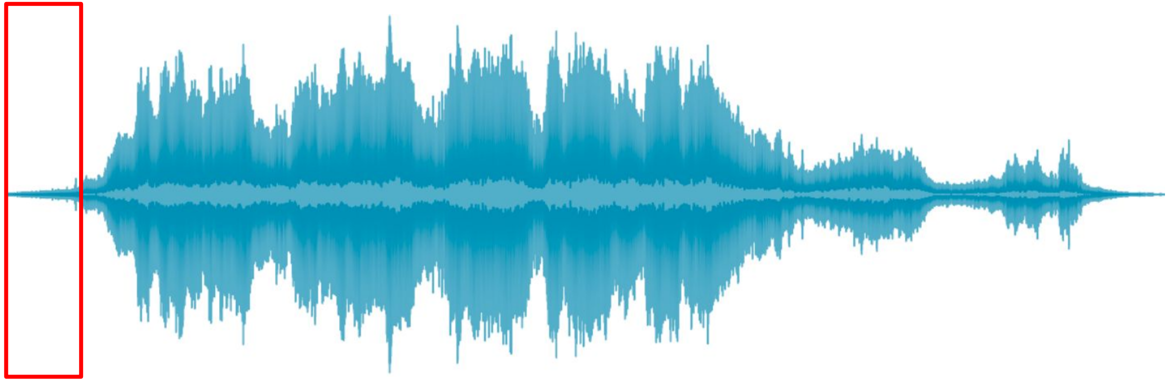
WE DON'T KNOW WHEN


CONSIDER SMALL SEGMENTS OF THE SIGNAL
APPLY FFT LOCALLY
imgflip.com

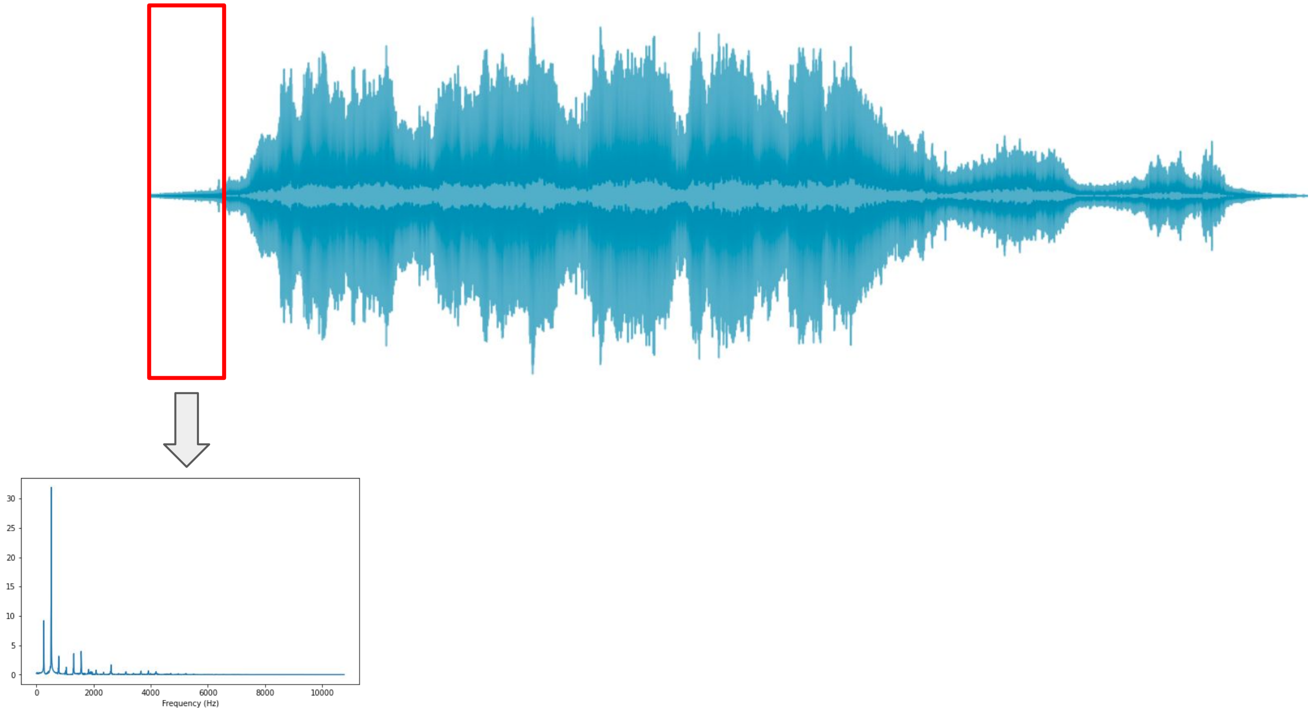
STFT intuition



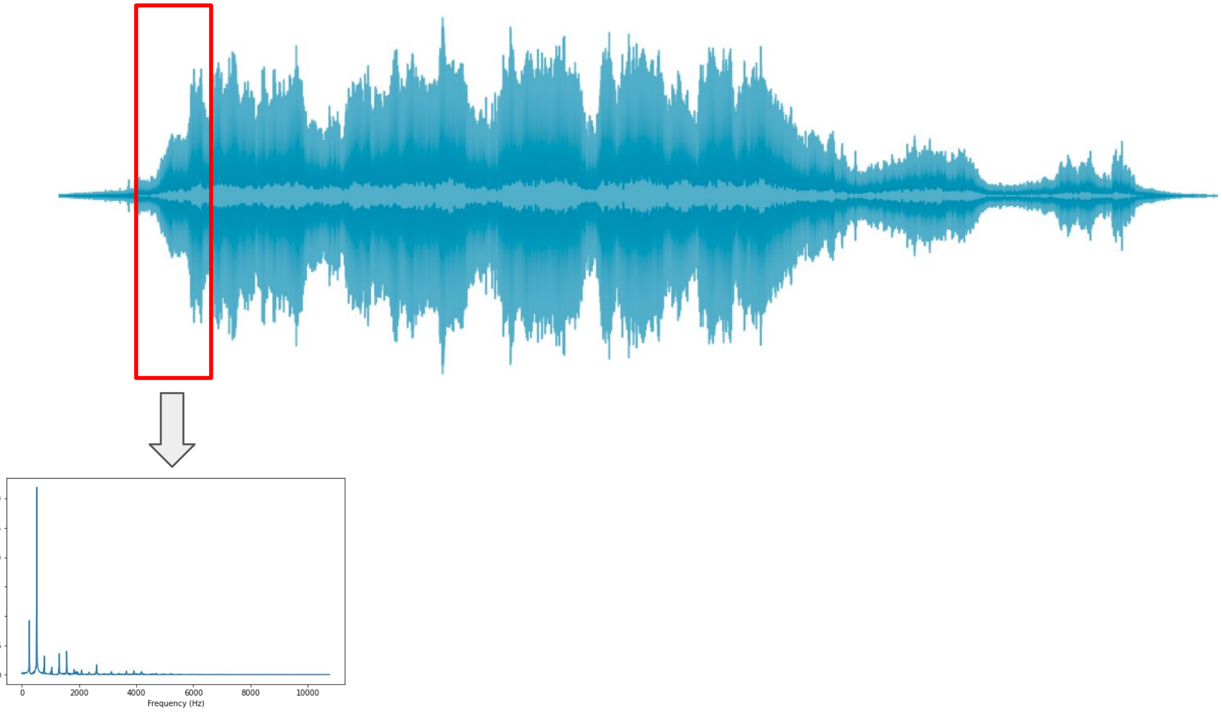
STFT intuition



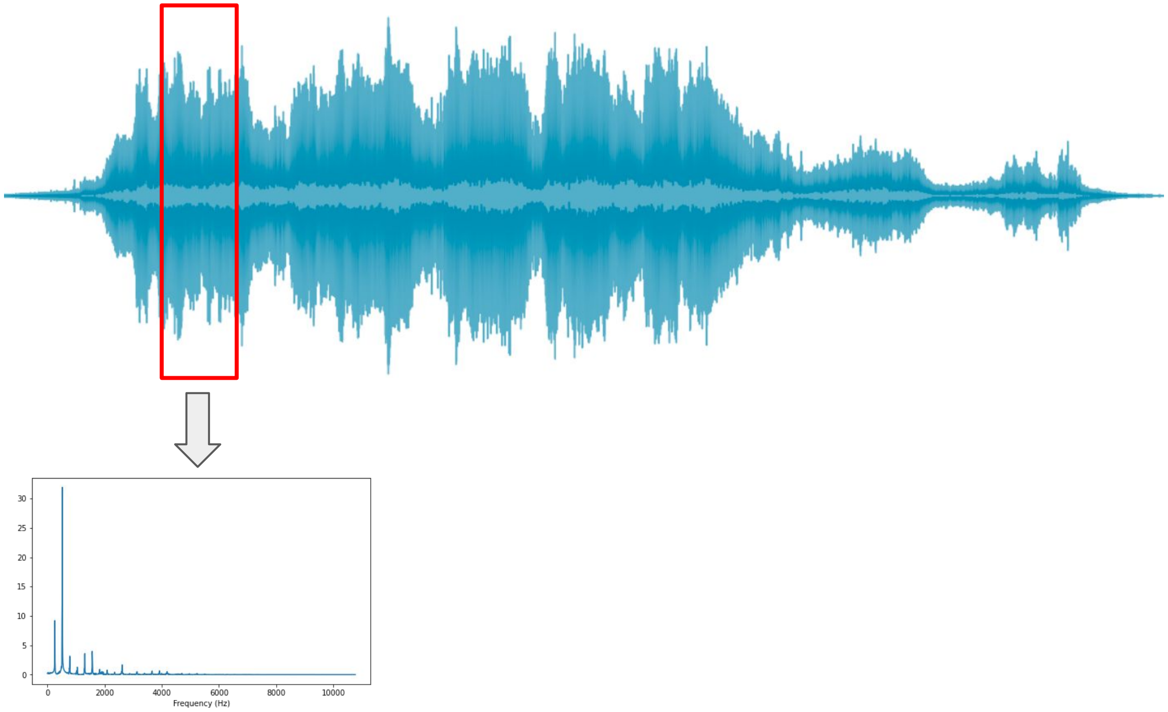
STFT intuition



STFT intuition



STFT intuition



Windowing

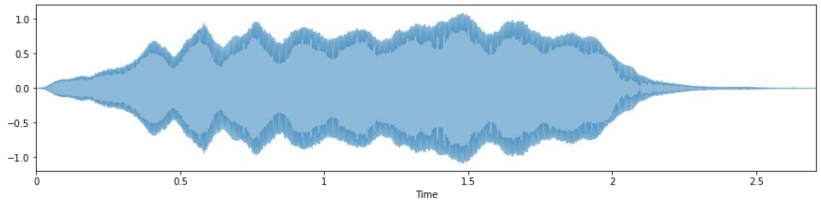
- Apply windowing function to signal

Windowing

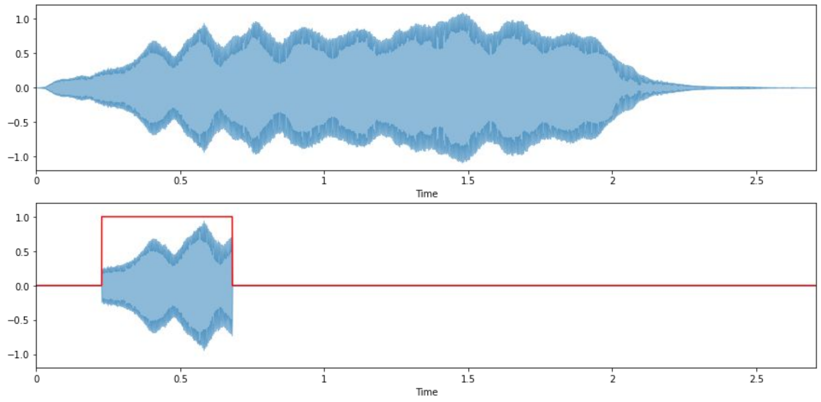
- Apply windowing function to signal

$$x_w(k) = x(k) \cdot w(k)$$

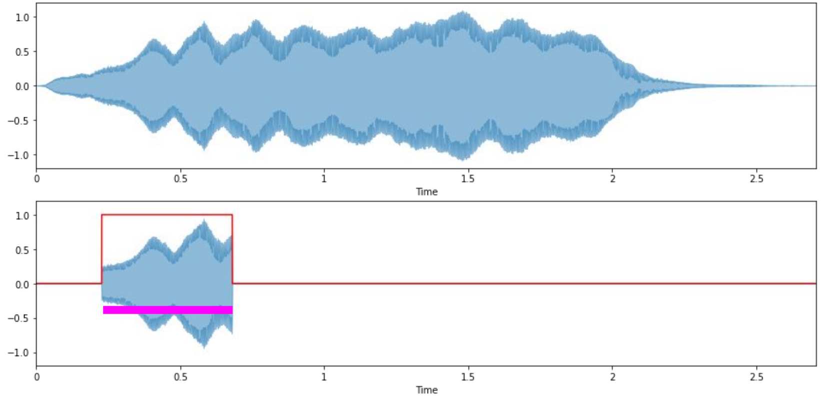
Windowing



Windowing

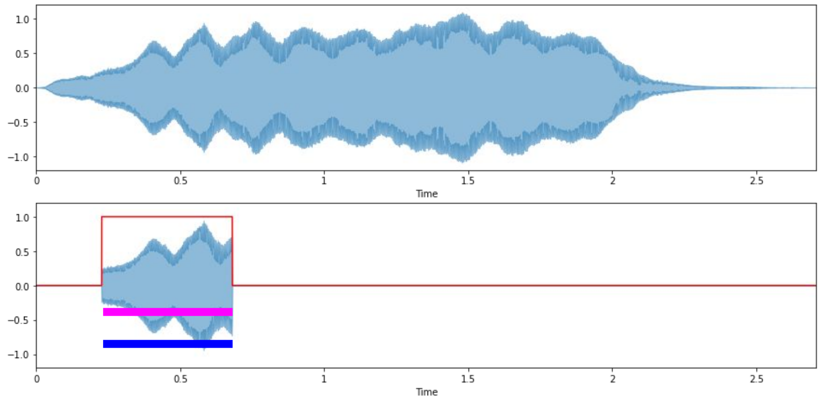


Windowing



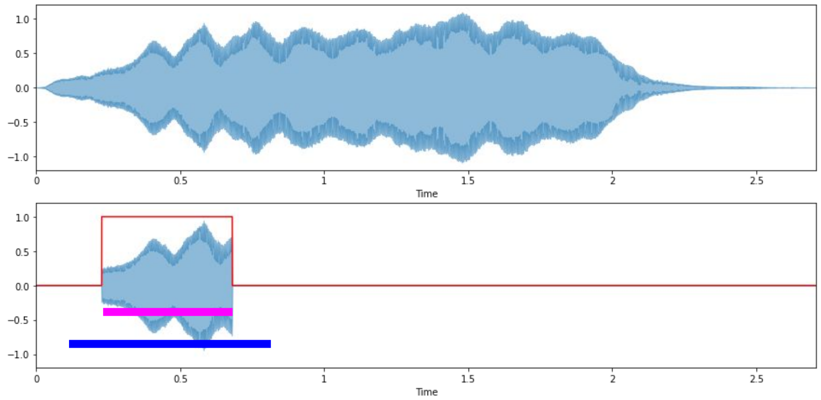
window size

Windowing



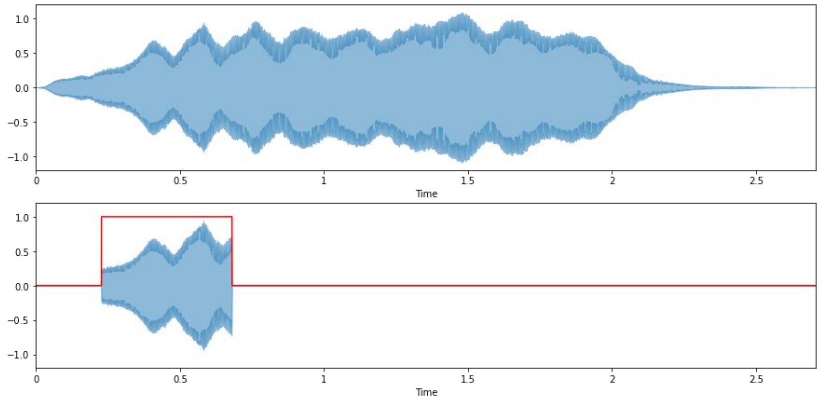
window size = frame size

Windowing

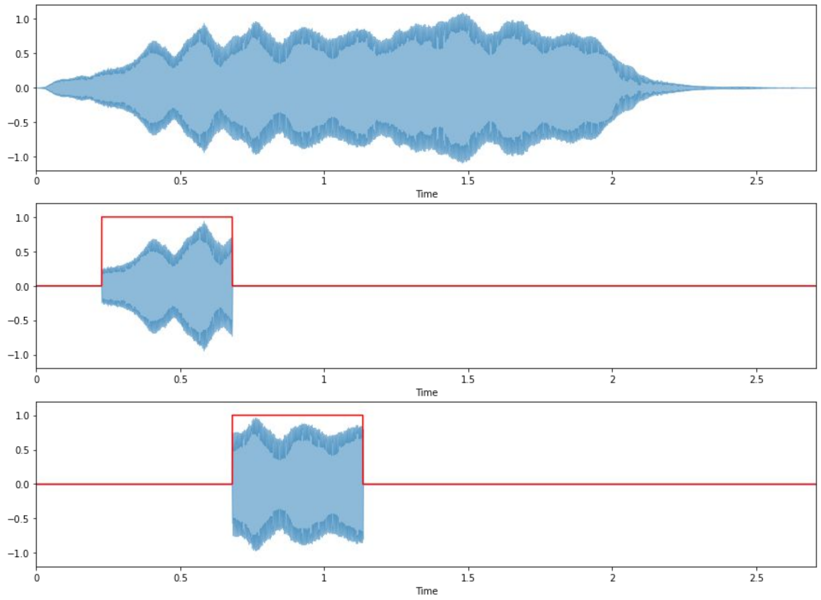


window size ≠ frame size

STFT

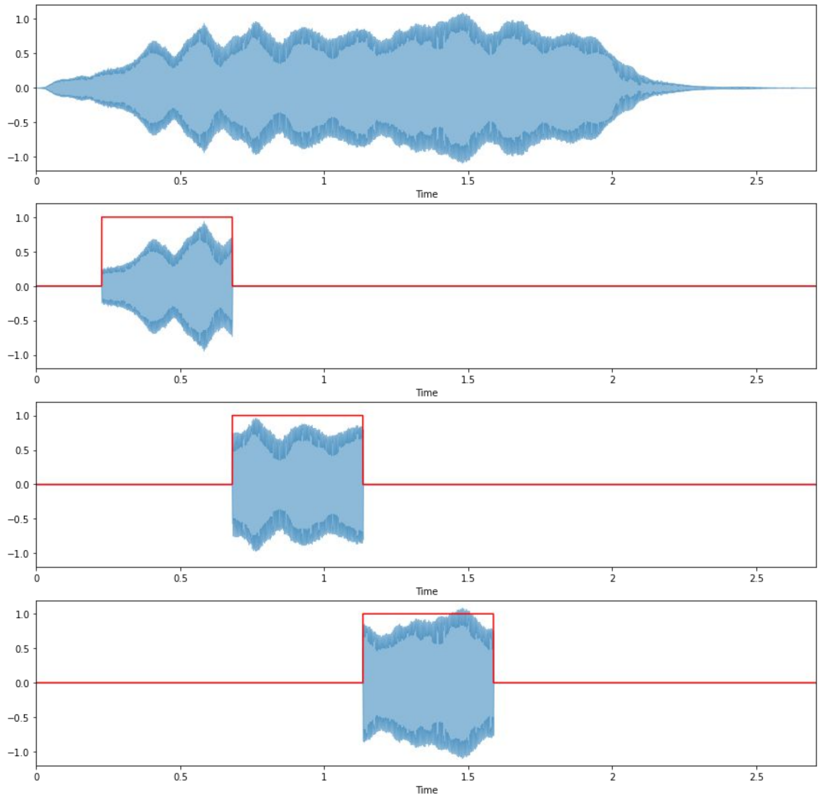


STFT

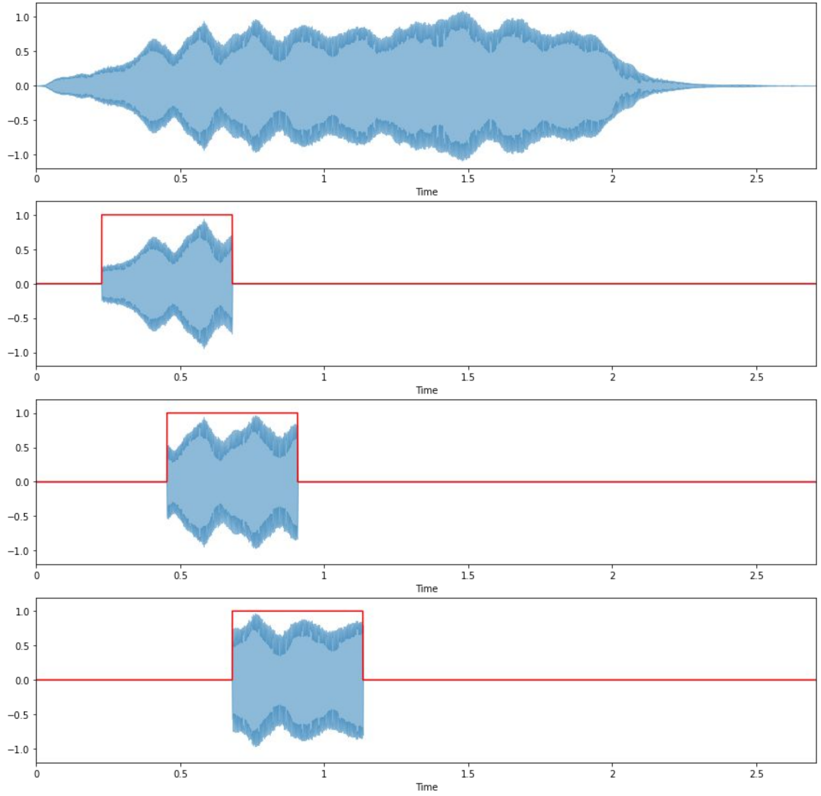


1.0
0.5
0.0
−0.5
−1.0
0
0.5
1
1.5
2
2.5
Time
1.0
0.5
0.0
−0.5
−1.0
0
0.5
1
1.5
2
2.5
Time
1.0
0.5
0.0
−0.5
−1.0
0
0.5
1
1.5
2
2.5
Time

STFT

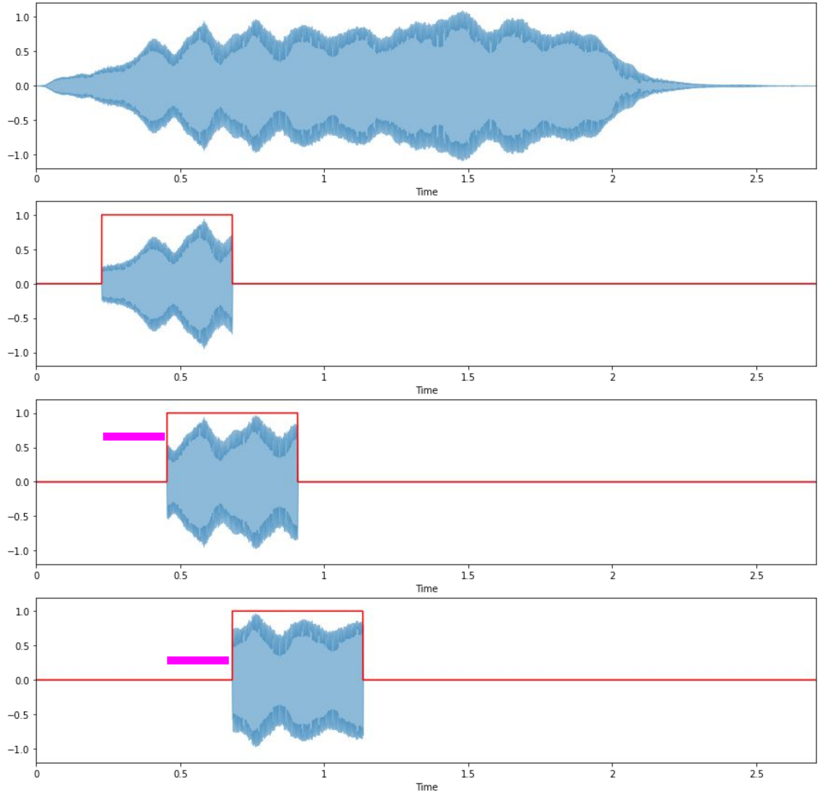


Overlapping frames



Overlapping frames

hop size (H)



From DFT to STFT

$$\hat{x}(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

$$S(m,k) = \sum_{n=0}^{N-1} x(n+mH) \cdot w(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

From DFT to STFT

$$\boxed{\hat{x}(k)} = \sum_{n=0}^{N-1} x(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

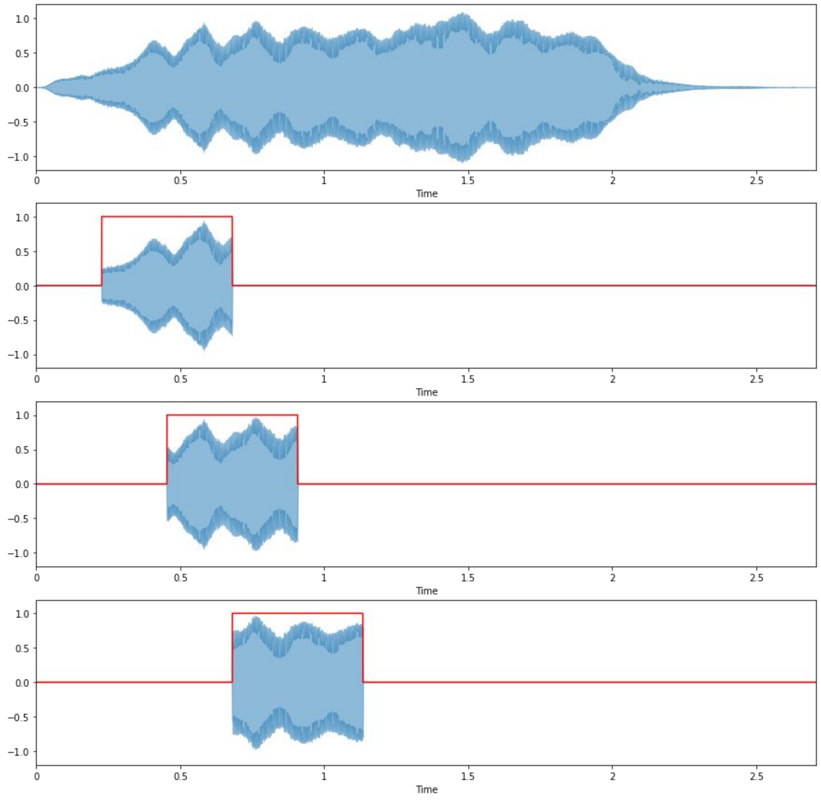
$$\boxed{S(m,k)} = \sum_{n=0}^{N-1} x(n + mH) \cdot w(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

From DFT to STFT

$$\boxed{\hat{x}(k)} = \sum_{n=0}^{N-1} x(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

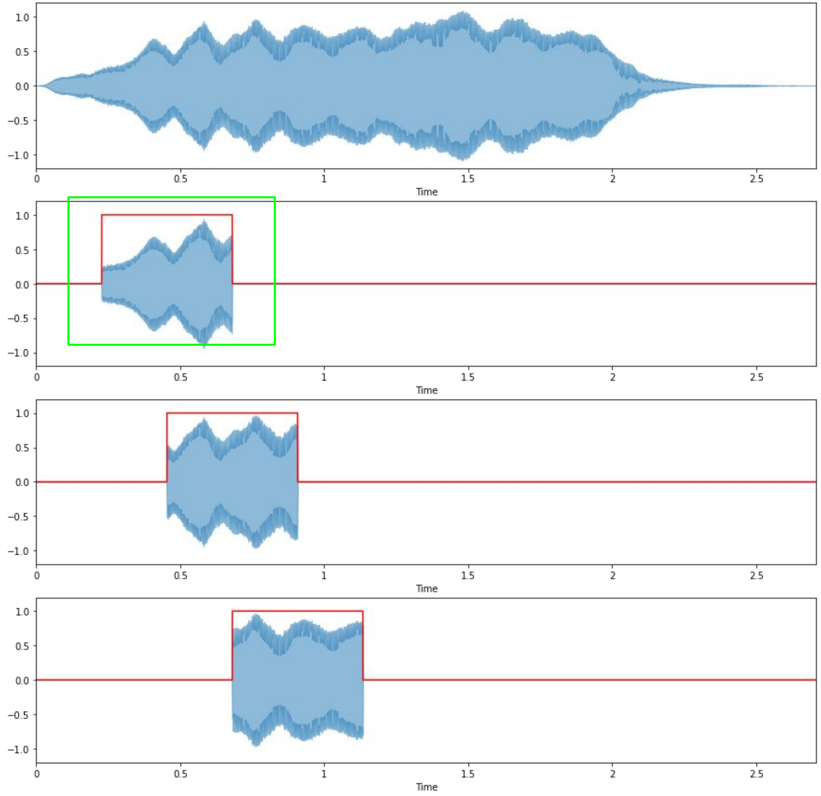
$$S(\boxed{m}, k) = \sum_{n=0}^{N-1} x(n + mH) \cdot w(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

From DFT to STFT



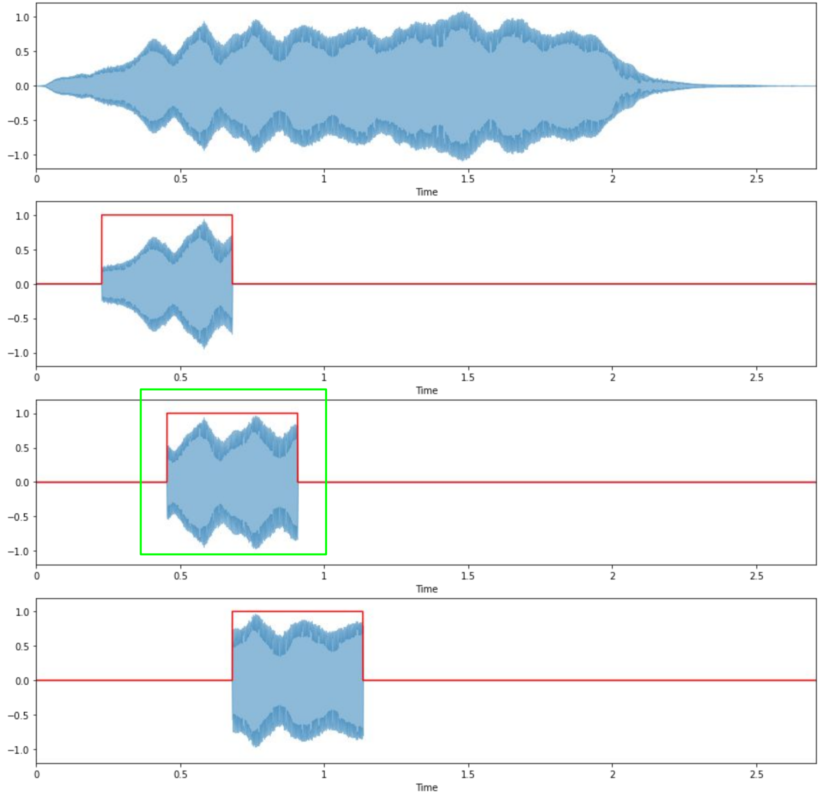
From DFT to STFT

m = 1

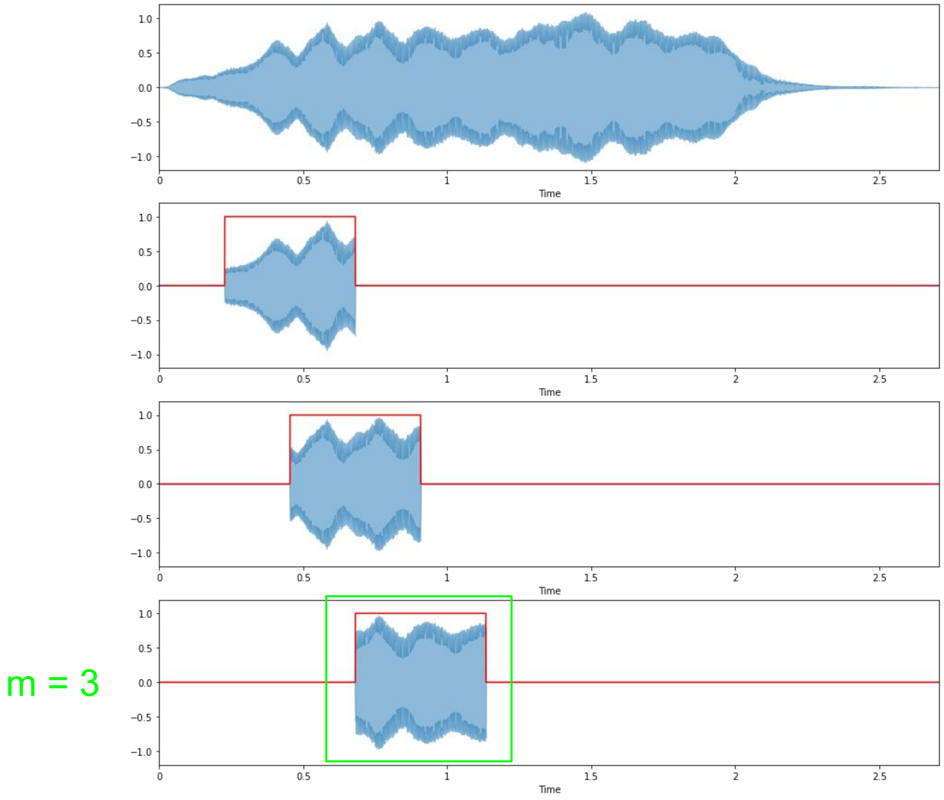


From DFT to STFT

m = 2



From DFT to STFT



From DFT to STFT

$$\hat{x}(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

$$\boxed{S(m,k)} = \sum_{n=0}^{N-1} x(n + mH) \cdot w(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

From DFT to STFT

$$\hat{x}(k) = \boxed{\sum_{n=0}^{N-1}} x(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

$$S(m,k) = \boxed{\sum_{n=0}^{N-1}} x(n + mH) \cdot w(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

From DFT to STFT

$$\hat{x}(k) = \sum_{n=0}^{N-1} \boxed{x(n)} \cdot e^{-i2\pi n \frac{k}{N}}$$

$$S(m,k) = \sum_{n=0}^{N-1} \boxed{x(n+mH)} \cdot w(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

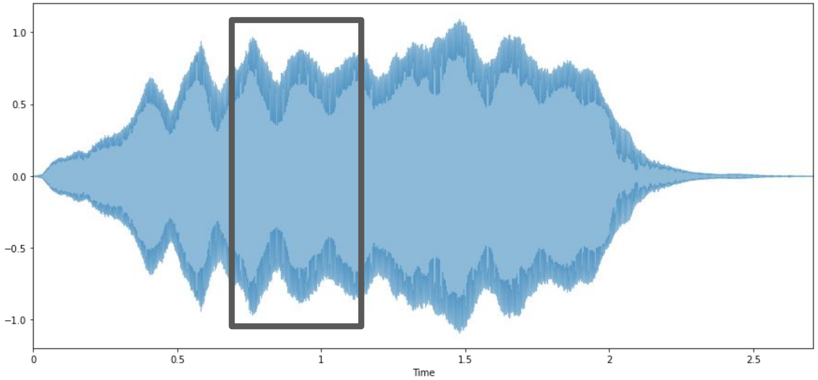
From DFT to STFT

$$\hat{x}(k) = \sum_{n=0}^{N-1} \boxed{x(n)} \cdot e^{-i2\pi n \frac{k}{N}}$$

$$S(m,k) = \sum_{n=0}^{N-1} \boxed{x(n + \boxed{mH})} \cdot w(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

Starting sample of
current frame

From DFT to STFT



From DFT to STFT

$$\hat{x}(k) = \sum_{n=0}^{N-1} \boxed{x(n)} \cdot e^{-i2\pi n \frac{k}{N}}$$

$$S(m,k) = \sum_{n=0}^{N-1} \boxed{x(n + \boxed{mH})} \cdot w(n) \cdot e^{-i2\pi n \frac{k}{N}}$$

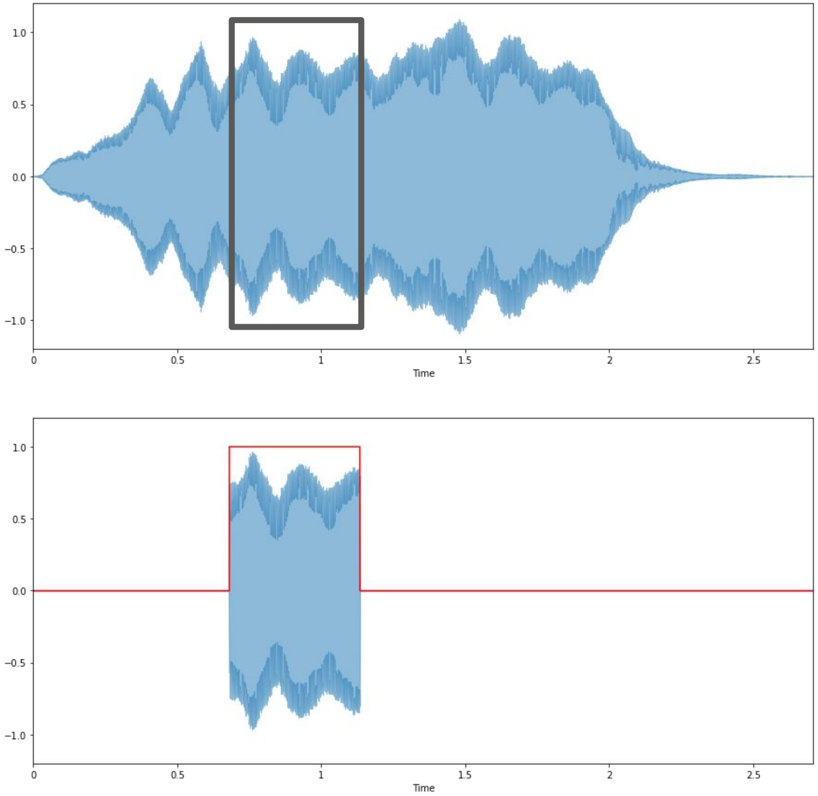
Starting sample of
current frame

From DFT to STFT

$$\hat{x}(k) = \sum_{n=0}^{N-1} \boxed{x(n)} \cdot e^{-i2\pi n \frac{k}{N}}$$

$$S(m,k) = \sum_{n=0}^{N-1} \boxed{x(n+mH) \cdot w(n)} \cdot e^{-i2\pi n \frac{k}{N}}$$

From DFT to STFT



From DFT to STFT

$$\hat{x}(k) = \sum_{n=0}^{N-1} x(n) \cdot \boxed{e^{-i2\pi n \frac{k}{N}}}$$

$$S(m,k) = \sum_{n=0}^{N-1} x(n+mH) \cdot w(n) \cdot \boxed{e^{-i2\pi n \frac{k}{N}}}$$

Outputs

- DFT
  - Spectral vector (# frequency bins)
  - N complex Fourier coefficients

Outputs

- DFT
  - Spectral vector (# frequency bins)
  - N complex Fourier coefficients
- STFT
  - Spectral matrix (# frequency bins, # frames)
  - Complex Fourier coefficients

Outputs

$$\text{\# frequency bins} = \frac{framesize}{2} + 1$$

Outputs

$$\text{\# frequency bins} = \frac{framesize}{2} + 1$$

$$\text{\# frames} = \frac{samples - framesize}{hopsize} + 1$$

Example STFT output

- Signal = 10K samples
- Frame size = 1000
- Hop size = 500

Example STFT output

- Signal = 10K samples
- Frame size = 1000
- Hop size = 500

frequency bins = 1000 / 2 + 1 = 501

Example STFT output

- Signal = 10K samples
- Frame size = 1000
- Hop size = 500

frequency bins = 1000 / 2 + 1 = 501 -> (0, sampling rate/2)

Example STFT output

- Signal = 10K samples
- Frame size = 1000
- Hop size = 500

frequency bins = 1000 / 2 + 1 = 501 -> (0, sampling rate/2)

frames = (10000 - 1000) / 500 + 1 = 19

Example STFT output

- Signal = 10K samples
- Frame size = 1000
- Hop size = 500

STFT -> (501, 19)

STFT parameters

- Frame size

STFT parameters

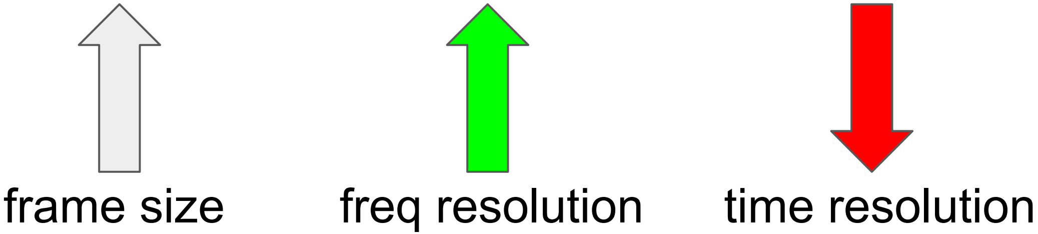
- Frame size

512, 1024, 2048, 4096, 8192

Time / frequency trade off

frame size

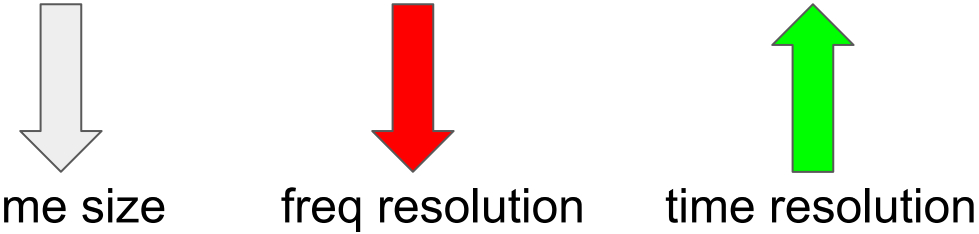
Time / frequency trade off



Time / frequency trade off

frame size

Time / frequency trade off



STFT parameters

- Frame size
- Hop size

STFT parameters

- Frame size
- Hop size

256, 512, 1024, 2048, 4096

STFT parameters

- Frame size
- Hop size

256, 512, 1024, 2048, 4096

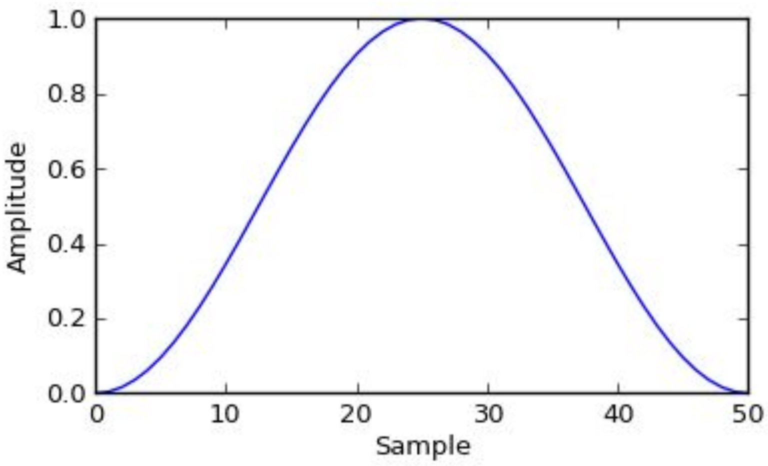
½ K, ¼ K, ⅛ K

STFT parameters

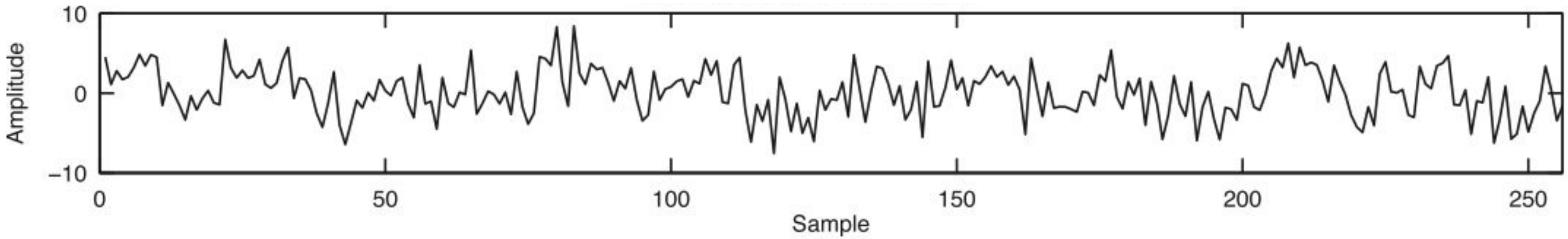
- Frame size
- Hop size
- Windowing function

Hann window

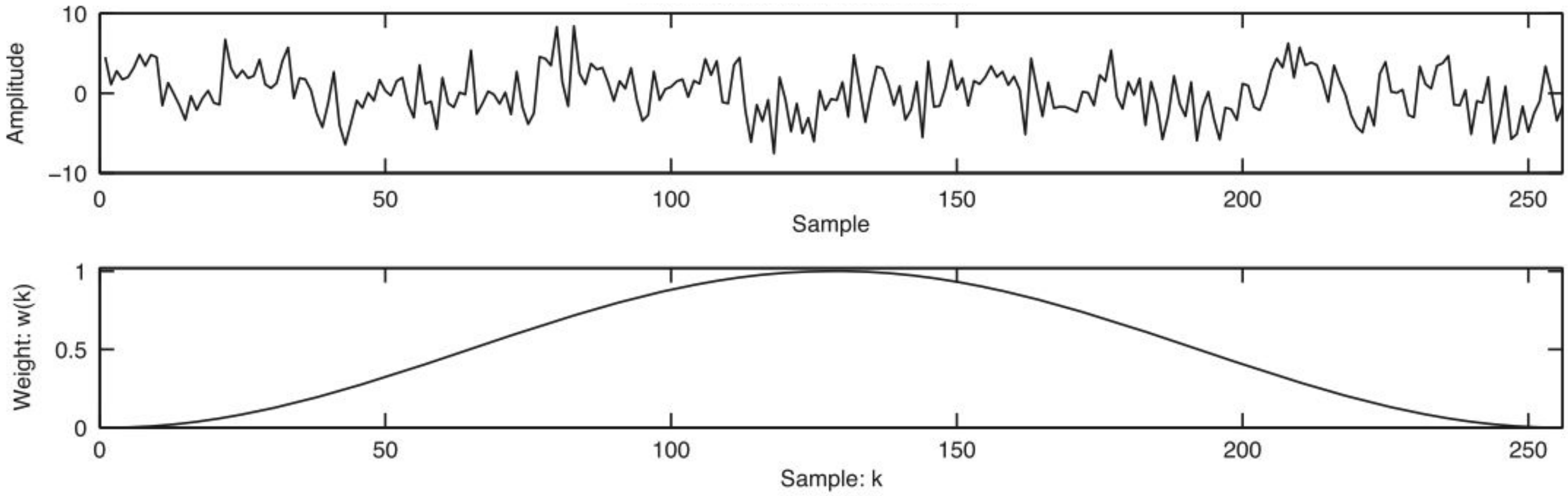
$$w(k) = 0.5 \cdot (1 - \cos(\frac{2\pi k}{K-1})), k = 1...K$$



Hann window

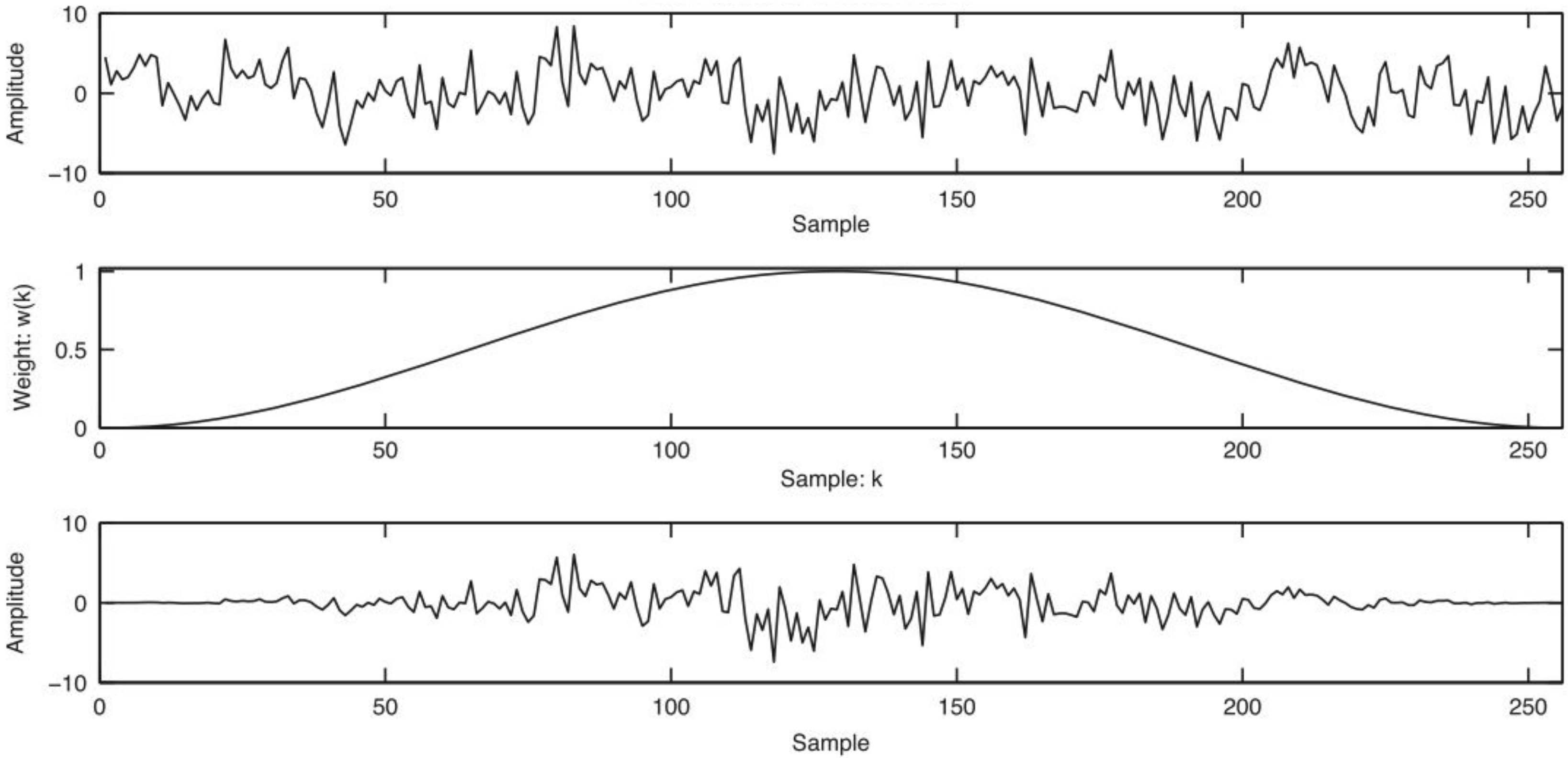


Hann window



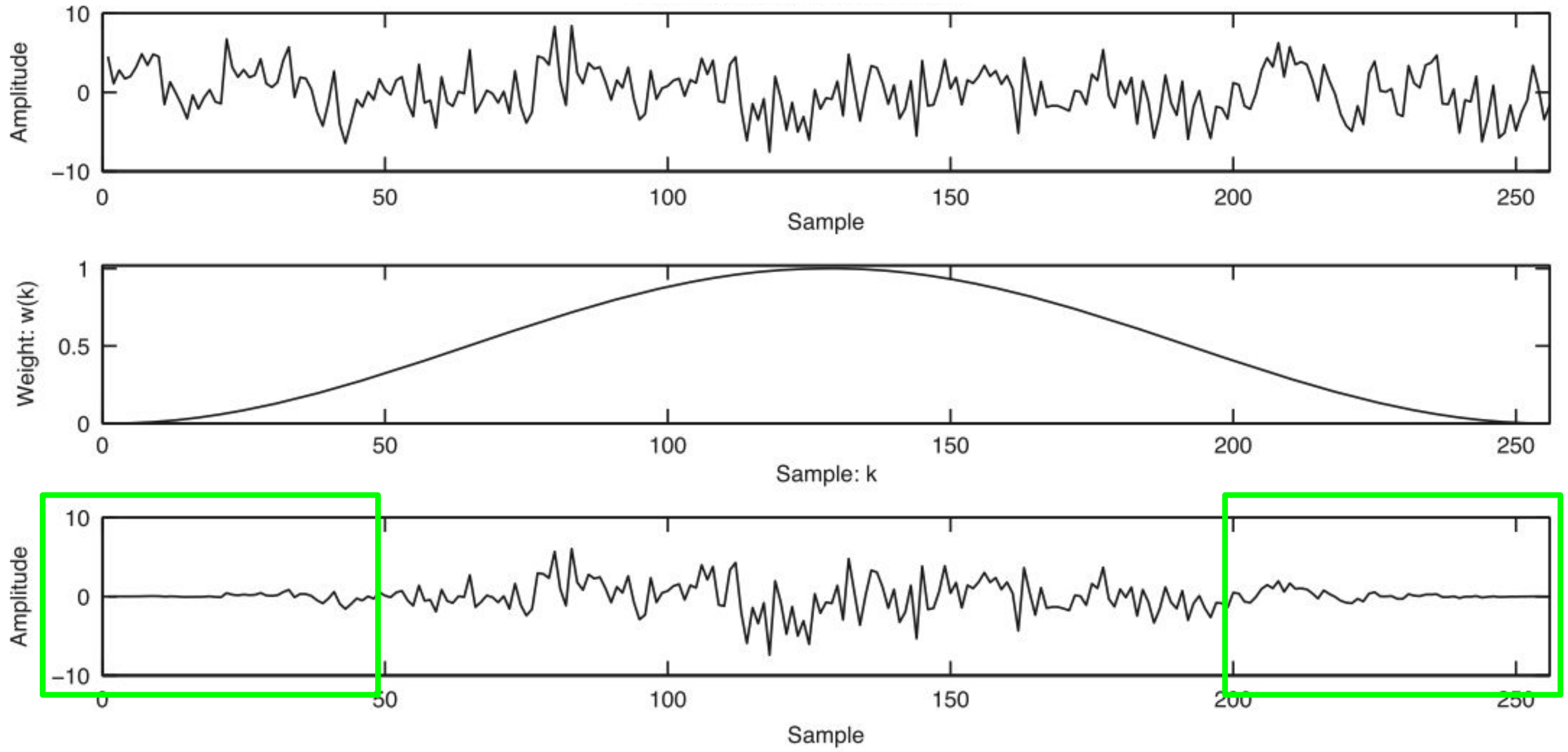
10
0
−10
Amplitude
0
50
100
150
200
250
Sample
1
0.5
0
Weight: w(k)
0
50
100
150
200
250
Sample: k

Hann window



10
0
−10
Amplitude
0
50
100
150
200
250
Sample
1
0.5
0
Weight: w(k)
Sample: k
Amplitude
Sample

Hann window



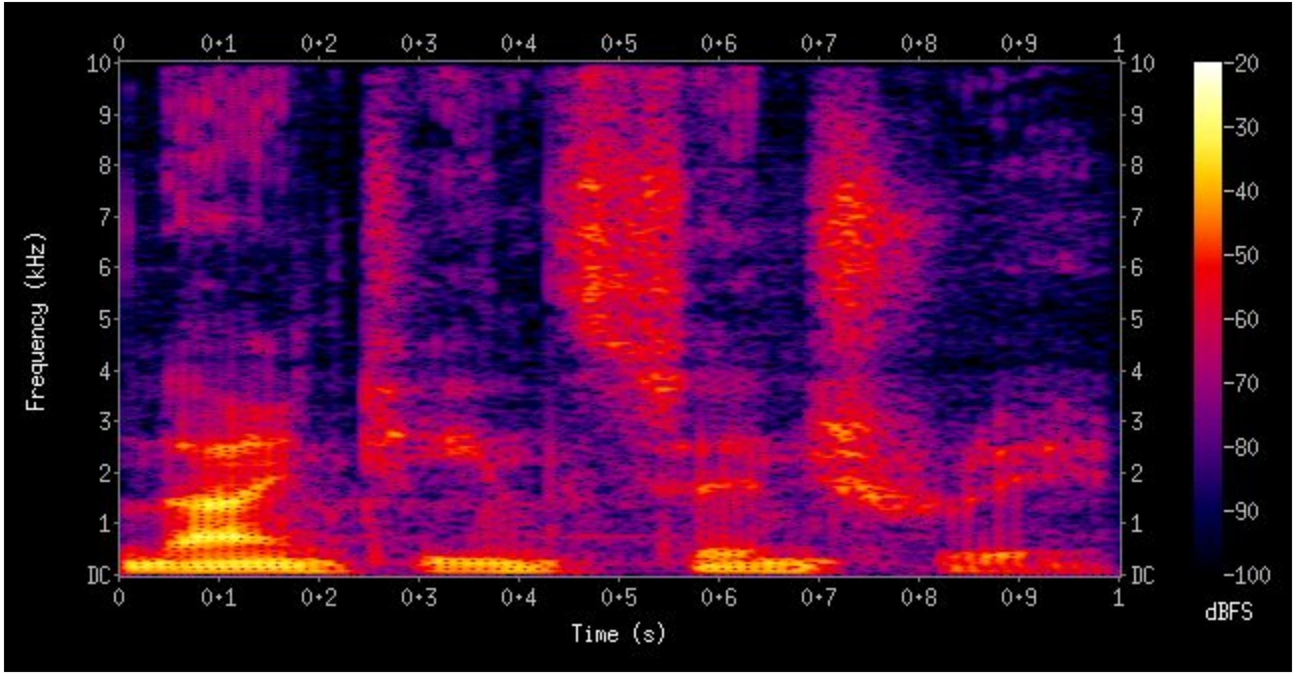
10
0
−10
Amplitude
0
50
100
150
200
250
Sample
1
0.5
0
Weight: w(k)
0
50
100
150
200
250
Sample: k
10
0
−10
Amplitude
0
50
100
150
200
250
Sample

Visualising sound

Visualising sound

$$Y(m,k) = |S(m,k)|^2$$

Spectrogram



What’s up next?

- Extract spectrograms with Librosa
- Discuss different flavours of spectrograms
- Examine different audio data