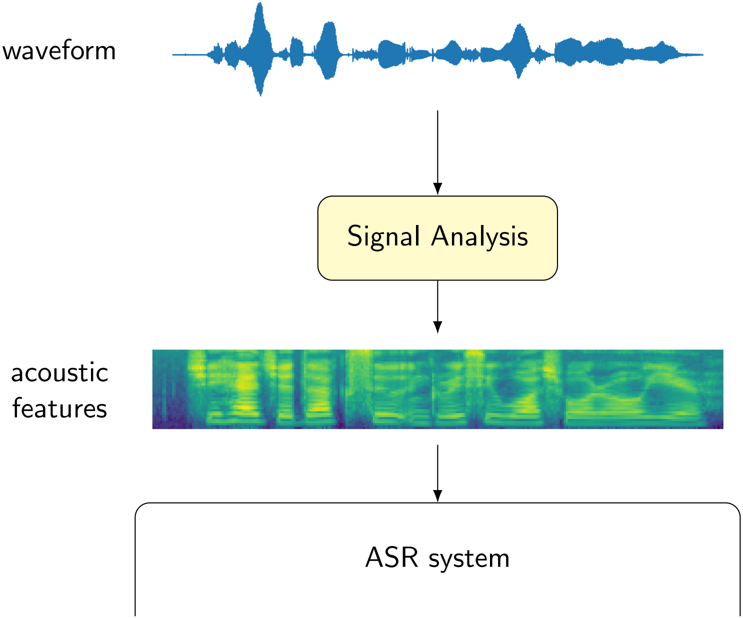
Speech Signal Analysis 1

Hao Tang

Automatic Speech Recognition—ASR Lecture 2
19 January 2023


waveform
Signal Analysis
acoustic features
ASR system

Outline

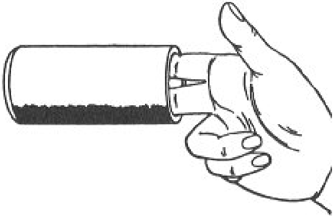
- Waveforms
  - Dithering
  - Removing DC offset
  - Pre-emphasis
- Spectrograms
  - Discrete Fourier transform (DFT)
  - Linearity and the shift theorem
  - Short-time Fourier transform
  - Windowing

waveform

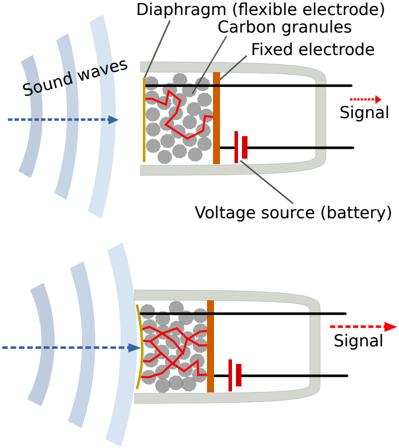
- Speech is part of sound waves.
- If we want to study speech, we need to be able to record, replay, and visualize speech.

Phonautograph (1857)

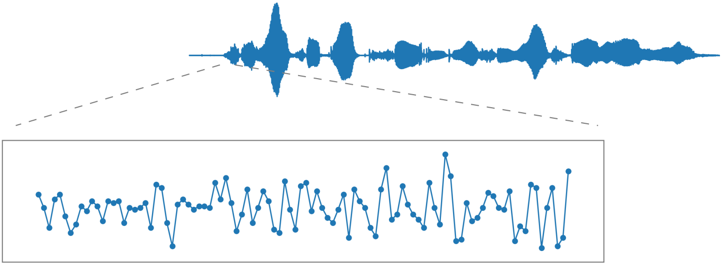
Phonograph (1877)



Carbon Microphone (1877)

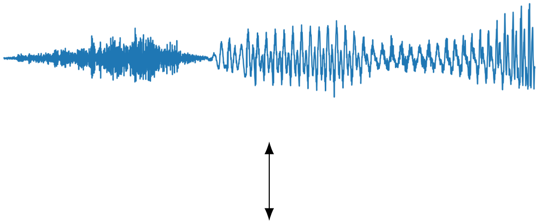


Wave Samples



- Sound waves are sampled and quantized.
- The typical sampling rate is 16,000 Hz. Each sample is typically a 16-bit integer.
- We will use $x[t]$ to denote the t-th sample in the signal x.

Line Plots and Vectors



$$\begin{bmatrix} -0.53 & -0.32 & 0.02 & 0.44 & \cdots & 0.18 \end{bmatrix}$$

Common Preprocessing in the Time Domain

- Dithering

$$y[t] = x[t] + \epsilon \qquad \epsilon \sim \mathcal{N}(0, 1)$$

  - Add a little Gaussian noise to the signal.
  - Avoid the signal being zeros, since we will be taking logarithm at some point.

- Removing DC offset

$$y[t] = x[t] - \frac{1}{T} \sum_{i=1}^{T} x[i]$$

  - Ensure that the signal has mean zero.
  - Most processing assumes the signal to have zero mean.

Common Preprocessing in the Time Domain

- Pre-emphasis

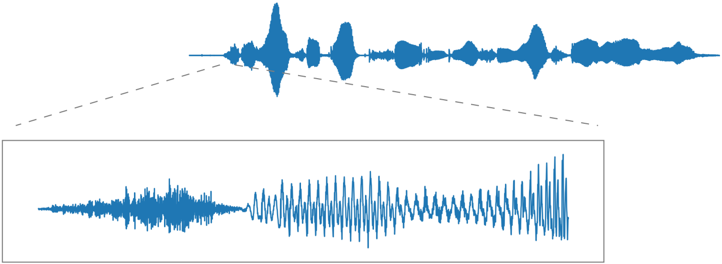
$$y[t] = x[t] - 0.97 \cdot x[t-1]$$

  - Emphasize the high-frequency components.
  - We will come back to this after we talked about frequency analysis.

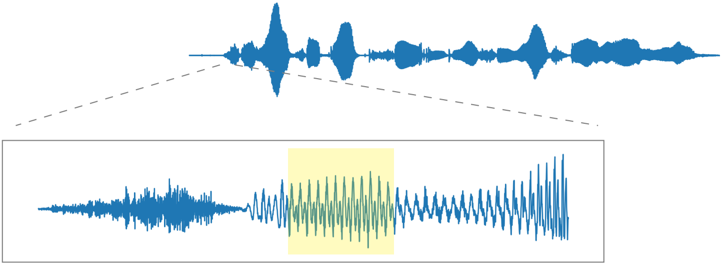
Ohm's Acoustic Law (1843)

If you hear a pitch of a certain frequency, then there must be energy of that frequency present in the sound wave.

Periodicity in Speech



Periodicity in Speech



Discrete Fourier Transform

$$X[k] = \sum_{t=0}^{T-1} x[t] e^{-i2\pi tk/T} \qquad \text{for } k = 0, \dots, T-1, \text{ and } i = \sqrt{-1}$$

Discrete Fourier Transform

$$X[k] = \sum_{t=0}^{T-1} x[t] e^{-i2\pi tk/T} \qquad \text{for } k = 0, \dots, T-1, \text{ and } i = \sqrt{-1}$$

$$X[k] = \begin{bmatrix} e^{i2\pi k \cdot 0/T} & e^{i2\pi k \cdot 1/T} & \cdots & e^{i2\pi k \cdot (T-1)/T} \end{bmatrix}^* \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[T-1] \end{bmatrix}$$

Discrete Fourier Transform

$$X[k] = \sum_{t=0}^{T-1} x[t] e^{-i2\pi tk/T} \qquad \text{for } k = 0, \dots, T-1, \text{ and } i = \sqrt{-1}$$

$$X[k] = \begin{bmatrix} e^{i2\pi k \cdot 0/T} & e^{i2\pi k \cdot 1/T} & \cdots & e^{i2\pi k \cdot (T-1)/T} \end{bmatrix}^* \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[T-1] \end{bmatrix}$$

$$v_k \triangleq \begin{bmatrix} e^{i2\pi k \cdot 0/T} & e^{i2\pi k \cdot 1/T} & \cdots & e^{i2\pi k \cdot (T-1)/T} \end{bmatrix}$$

$$e^{i\theta} = \cos\theta + i \sin\theta$$

Fourier Basis

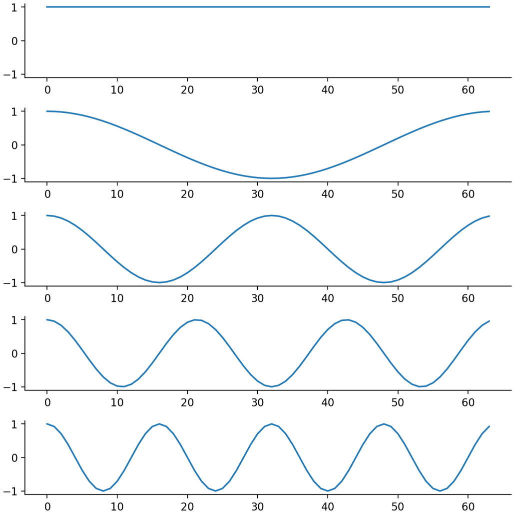
$\Re\{v_0\}$

$\Re\{v_1\}$

$\Re\{v_2\}$

$\Re\{v_3\}$

$\Re\{v_4\}$



Fourier Basis

- The larger the k, the higher the frequency.

$$v_k = \begin{bmatrix} e^{i2\pi k \cdot 0/T} & e^{i2\pi k \cdot 1/T} & \cdots & e^{i2\pi k \cdot (T-1)/T} \end{bmatrix}$$

Fourier Basis

- The larger the k, the higher the frequency.

$$v_k = \begin{bmatrix} e^{i2\pi k \cdot 0/T} & e^{i2\pi k \cdot 1/T} & \cdots & e^{i2\pi k \cdot (T-1)/T} \end{bmatrix}$$

- The set $\{v_0/T, v_1/T, \dots, v_{T-1}/T\}$ is an orthonormal basis.

$$v_m^* v_n = \begin{cases} 0 & \text{if } m \neq n \\ T & \text{if } m = n \end{cases}$$

Fourier Basis

- The larger the k, the higher the frequency.

$$v_k = \begin{bmatrix} e^{i2\pi k \cdot 0/T} & e^{i2\pi k \cdot 1/T} & \dots & e^{i2\pi k \cdot (T-1)/T} \end{bmatrix}$$

- The set $\{v_0/T, v_1/T, \dots, v_{T-1}/T\}$ is an orthonormal basis.

$$v_m^* v_n = \begin{cases} 0 & \text{if } m \neq n \\ T & \text{if } m = n \end{cases}$$

- Fourier transform is a change of coordinates.

Discrete Fourier Transform

$$X[k] = \sum_{t=0}^{T-1} x[t] e^{-i2\pi tk/T} = v_k^* x$$

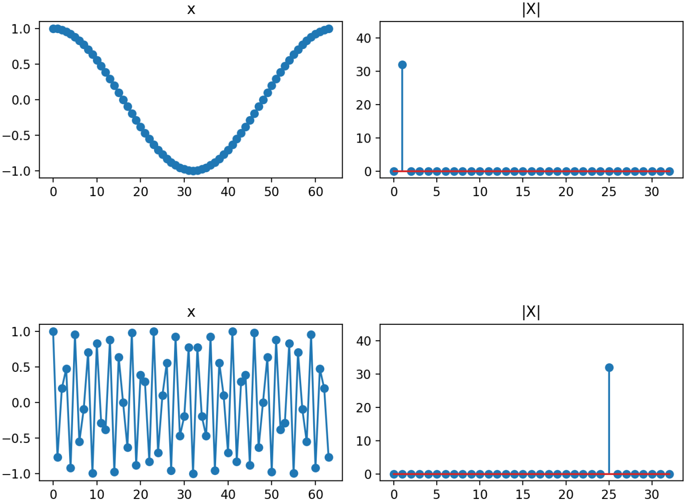
- $X[k]$ is a complex number.
- $X[k]$ is a (complex) dot product of a complex sinusoid v_k and the signal x.
- $X[k]$ tells us how similar x is to v_k.
- The large k's in X are high-frequency components, while the small k's in X are low-frequency components.

Discrete Fourier Transform

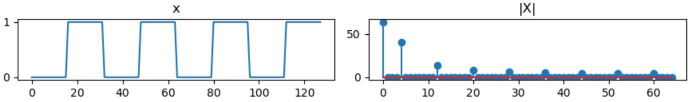
$$X = \mathcal{F}\{x\}$$

- DFT decomposes a signal into frequency components.
- X is also called the spectrum of x.

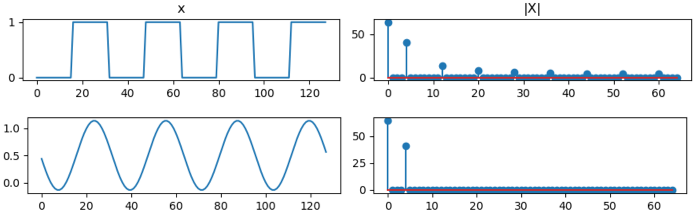
Discrete Fourier Transform



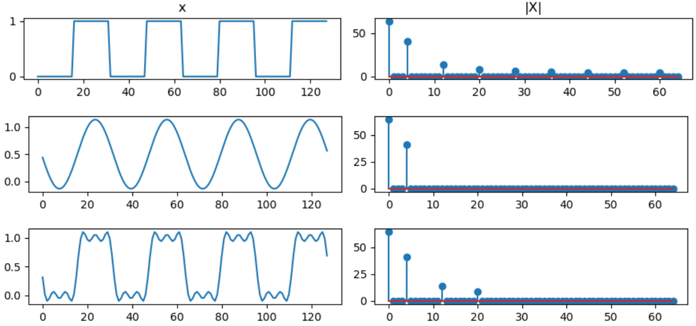
Discrete Fourier Transform



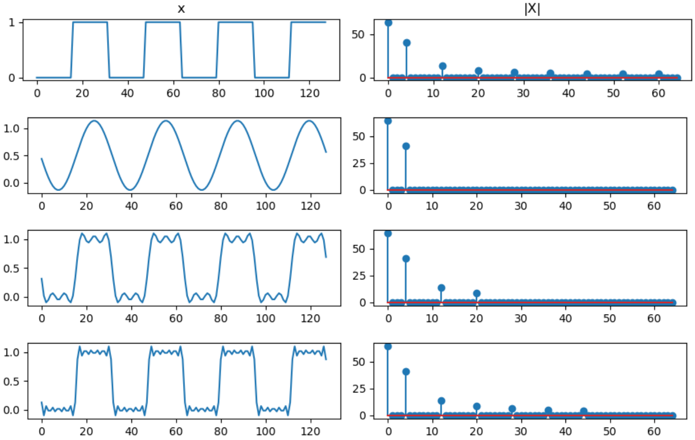
Discrete Fourier Transform



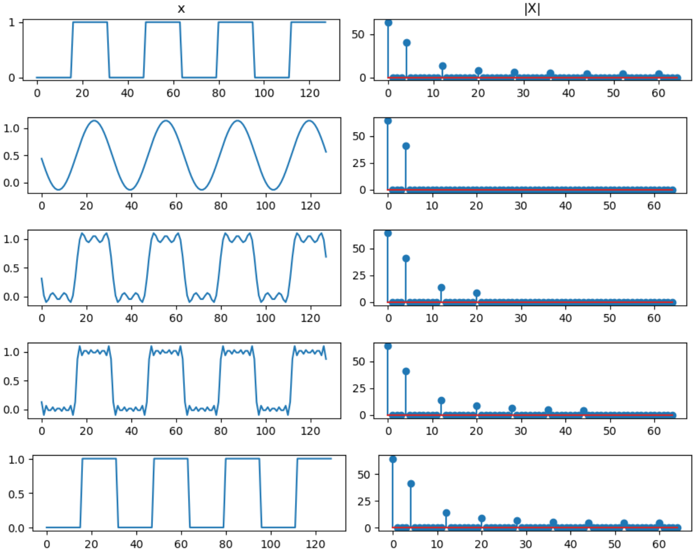
Discrete Fourier Transform



Discrete Fourier Transform



Discrete Fourier Transform



Properties of DFT

- Linearity

$$\mathcal{F}\{a_1 x_1 + a_2 x_2\} = a_1 \mathcal{F}\{x_1\} + a_2 \mathcal{F}\{x_2\}$$

Properties of DFT

- Linearity

$$\mathcal{F}\{a_1 x_1 + a_2 x_2\} = a_1 \mathcal{F}\{x_1\} + a_2 \mathcal{F}\{x_2\}$$

- Shift Theorem

$$\text{If } y[t] = x[t-1], \text{ then } Y[k] = e^{i2\pi k/T} X[k].$$

Proof of the Shift Theorem

$$Y[k] = \sum_{t=0}^{T-1} y[t] e^{-i2\pi tk/T}$$

Proof of the Shift Theorem

$$
\begin{aligned}
Y[k] &= \sum_{t=0}^{T-1} y[t] e^{-i2\pi tk/T} \\
&= \sum_{t=0}^{T-1} x[t-1] e^{-i2\pi tk/T}
\end{aligned}
$$

Proof of the Shift Theorem

$$
\begin{aligned}
Y[k] &= \sum_{t=0}^{T-1} y[t] e^{-i2\pi tk/T} \\
&= \sum_{t=0}^{T-1} x[t-1] e^{-i2\pi tk/T} \\
&= e^{-i2\pi k/T} \sum_{t=0}^{T-1} x[t-1] e^{-i2\pi (t-1)k/T}
\end{aligned}
$$

Proof of the Shift Theorem

$$
\begin{aligned}
Y[k] &= \sum_{t=0}^{T-1} y[t] e^{-i2\pi tk/T} \\
&= \sum_{t=0}^{T-1} x[t-1] e^{-i2\pi tk/T} \\
&= e^{-i2\pi k/T} \sum_{t=0}^{T-1} x[t-1] e^{-i2\pi (t-1)k/T} \\
&= e^{-i2\pi k/T} X[k]
\end{aligned}
$$

Pre-emphasis

- Definition

$$y[t] = x[t] - 0.97 \cdot x[t-1]$$

Pre-emphasis

- Definition

$$y[t] = x[t] - 0.97 \cdot x[t-1]$$

- DFT of pre-emphasis

$$\begin{aligned} Y[k] &= X[k] - 0.97 \cdot e^{-i2\pi k/T} X[k] \\ &= (1 - 0.97 \cdot e^{-i2\pi k/T}) X[k] \end{aligned}$$

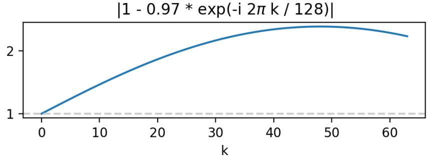
Pre-emphasis

- Definition

$$y[t] = x[t] - 0.97 \cdot x[t-1]$$

- DFT of pre-emphasis

$$\begin{aligned} Y[k] &= X[k] - 0.97 \cdot e^{-i2\pi k/T} X[k] \\ &= (1 - 0.97 \cdot e^{-i2\pi k/T}) X[k] \end{aligned}$$



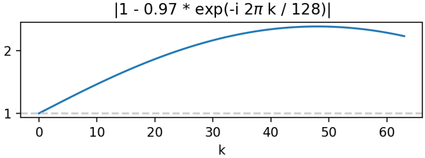
Pre-emphasis

- Definition

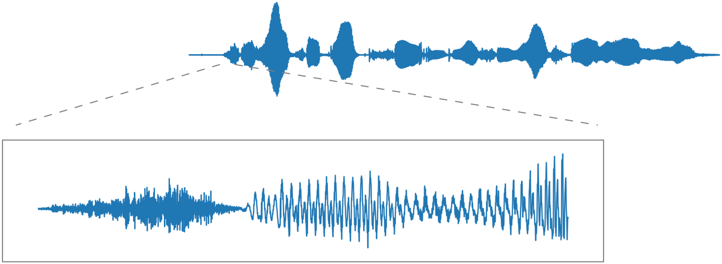
$$y[t] = x[t] - 0.97 \cdot x[t-1]$$

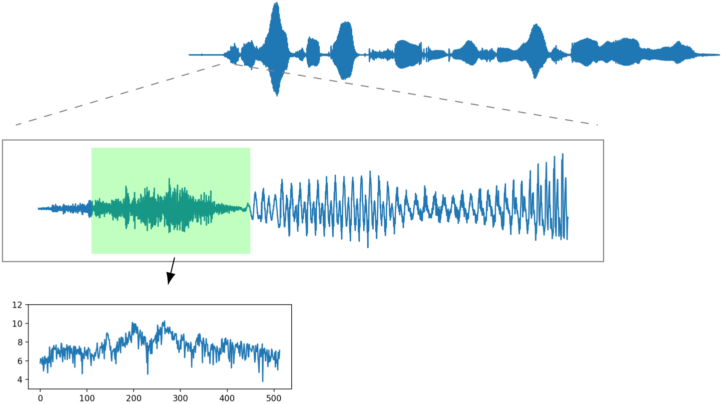
- DFT of pre-emphasis

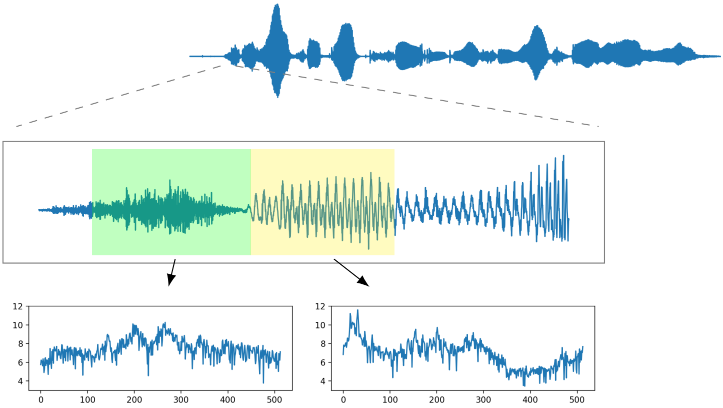
$$\begin{aligned} Y[k] &= X[k] - 0.97 \cdot e^{-i2\pi k/T} X[k] \\ &= (1 - 0.97 \cdot e^{-i2\pi k/T}) X[k] \end{aligned}$$

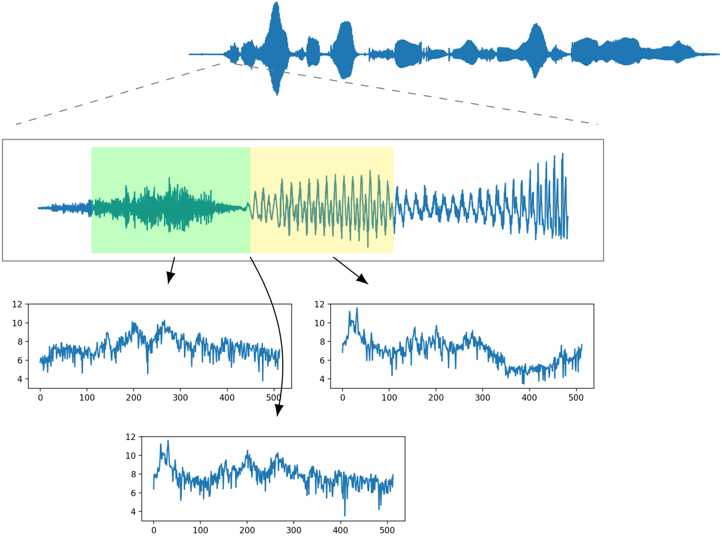


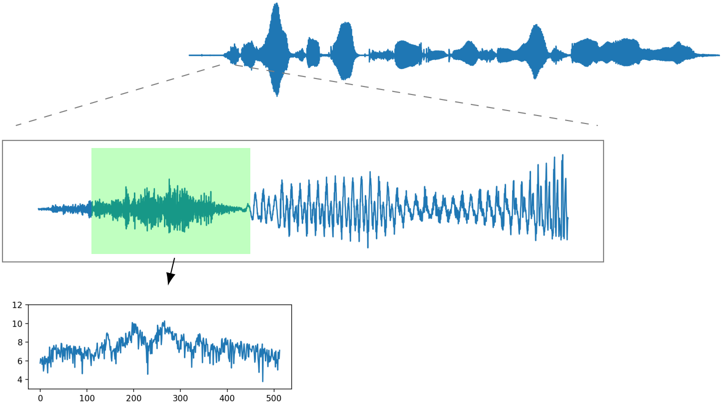
- In other words, pre-emphasis emphsizes the high-frequency region.

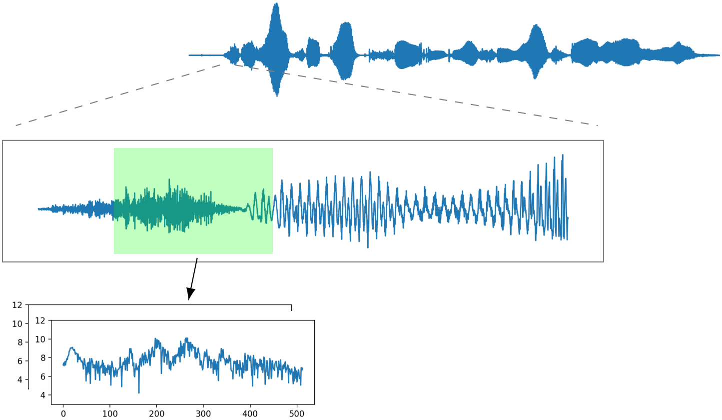


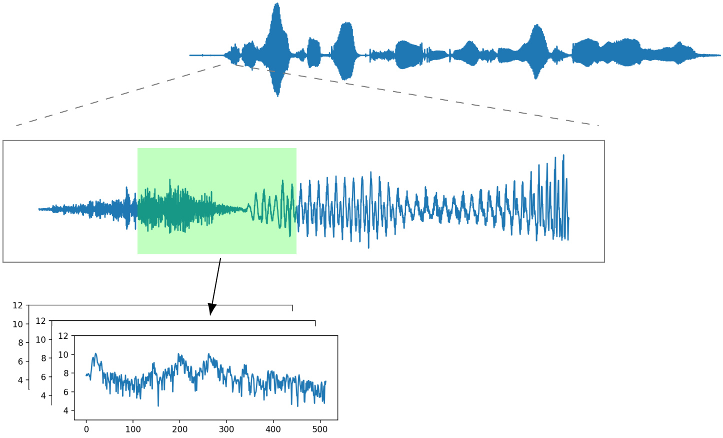

12
10
8
6
4
0
100
200
300
400
500

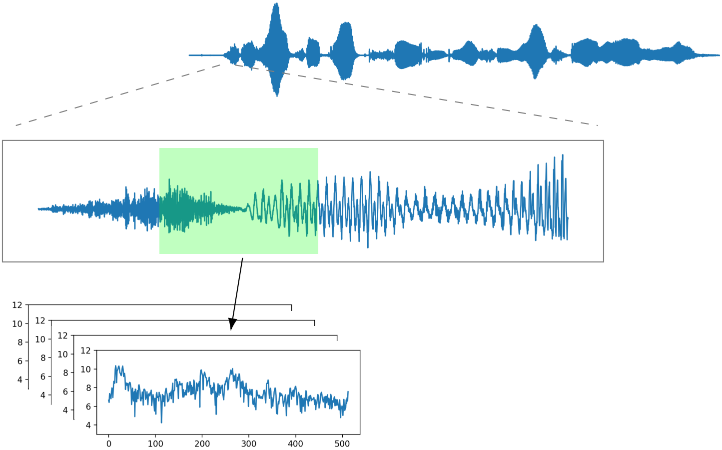

12
10
8
6
4
0
100
200
300
400
500

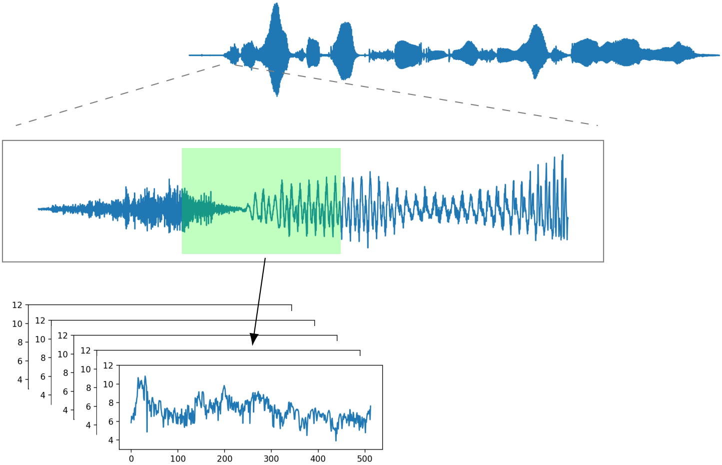

12
10
8
6
4
0
100
200
300
400
500

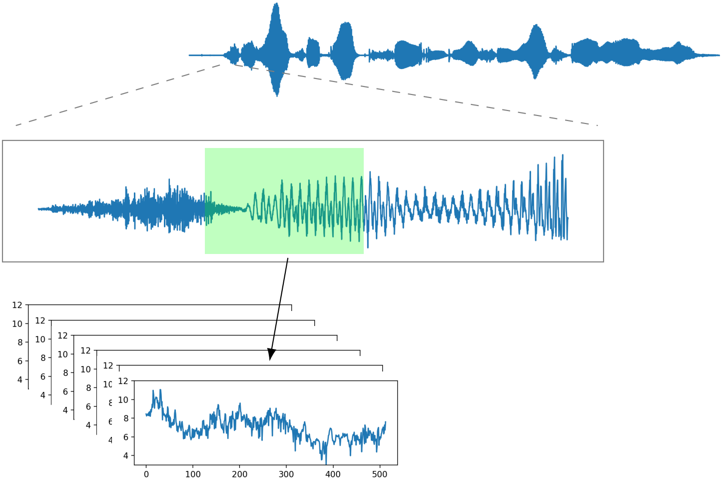

12
10
8
6
4
0
100
200
300
400
500

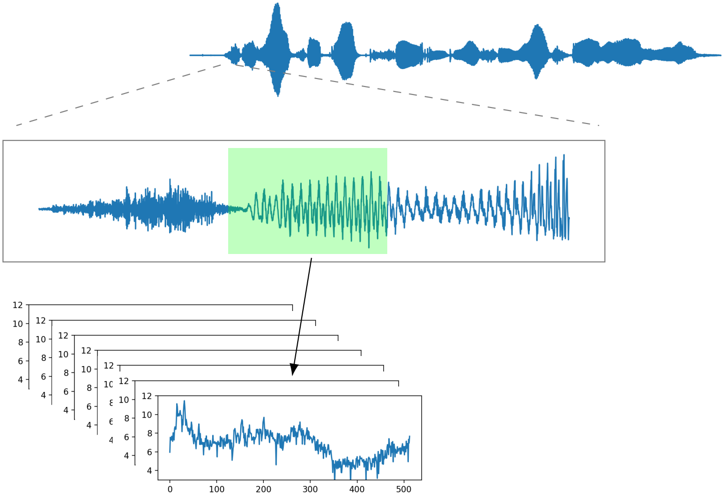

12
10
8
6
4
12
10
8
6
4
0
100
200
300
400
500

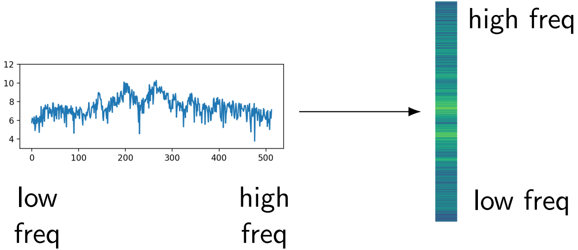

12
10
8
6
4
0
100
200
300
400
500

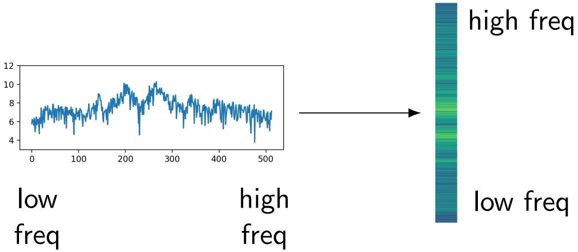

12
10
8
6
4
0
100
200
300
400
500

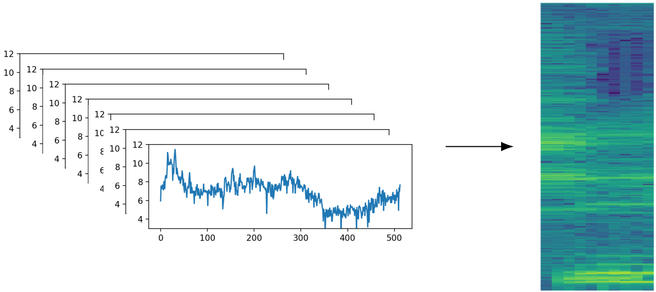

12
10
8
6
4
0
100
200
300
400
500


12
10
8
6
4
0
100
200
300
400
500

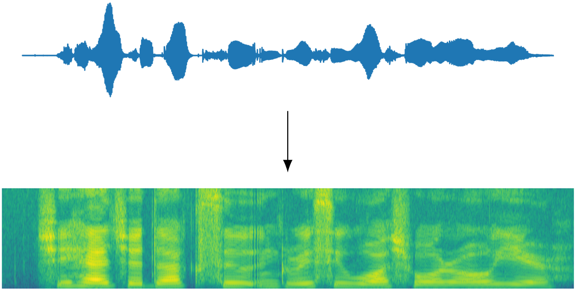

12
10
8
6
4
0
100
200
300
400
500


12
10
8
6
4
0
100
200
300
400
500
low freq
high freq
high freq
low freq

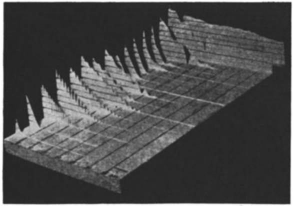

high freq
low freq
low
freq
high
freq

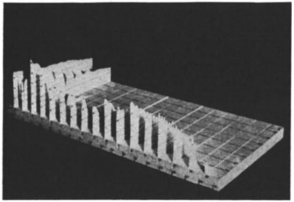


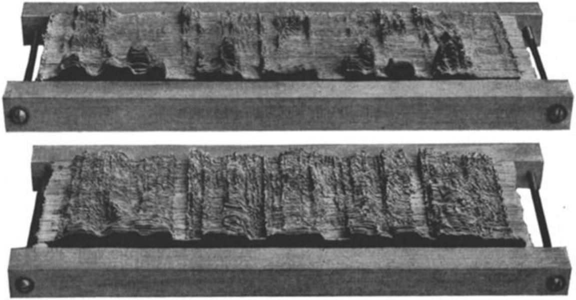
Short-Time Fourier Transform



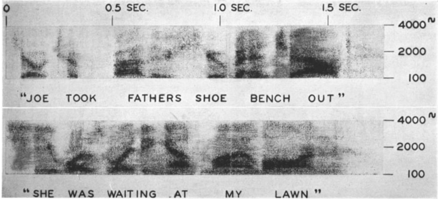
- Speech is non-stationary.
- Extract spectra with a sliding window, typically with a 25ms window size and a 10ms hop.
- Display the spectra as a heat map.

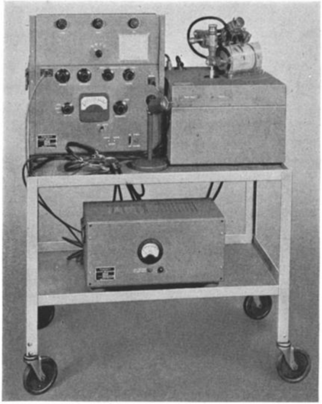


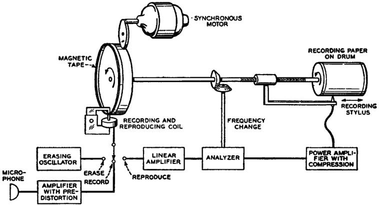




Sound Spectrograph (1946)

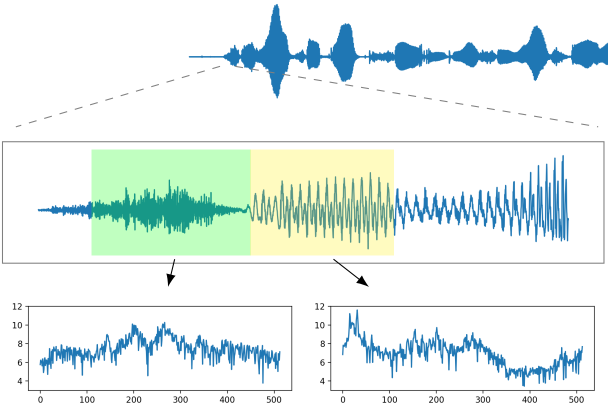




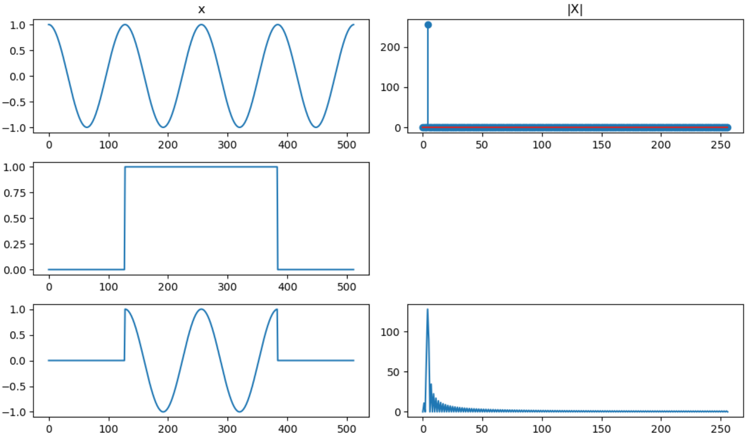


Fast Fourier Transform (1965)

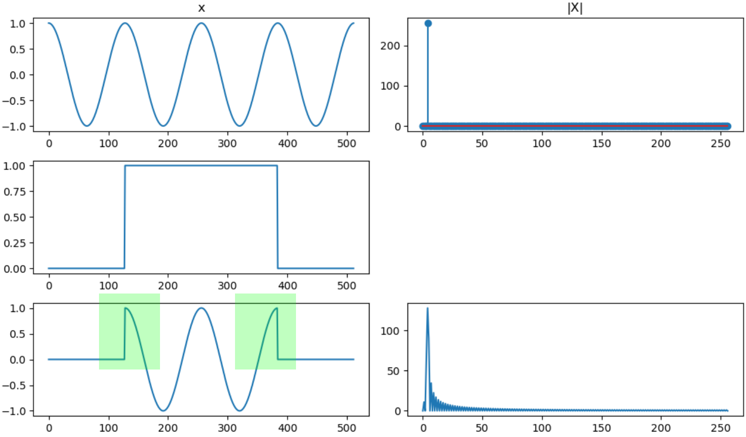
- The algorithm that we know of today was proposed in 1965.
- It was applied to speech on a computer around 1969.


12
10
8
6
4
0
100
200
300
400
500
12
10
8
6
4
0
100
200
300
400
500

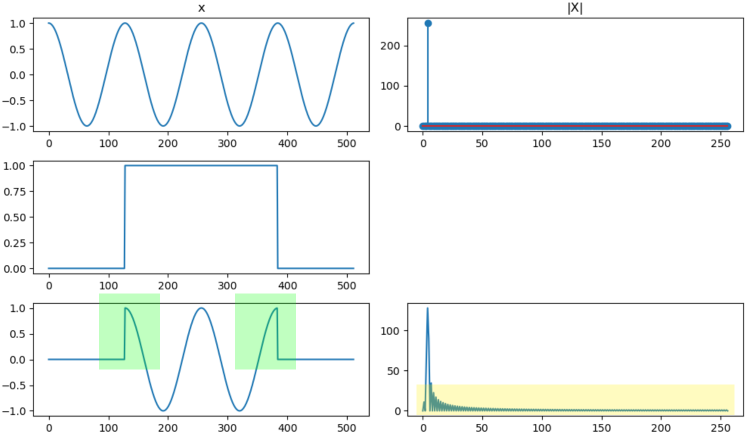
Windowing


x
|X|
1.0
0.5
0.0
−0.5
−1.0
0
100
200
300
400
500
200
100
0
50
150
250
1.00
0.75
0.50
0.25
0.00

Windowing

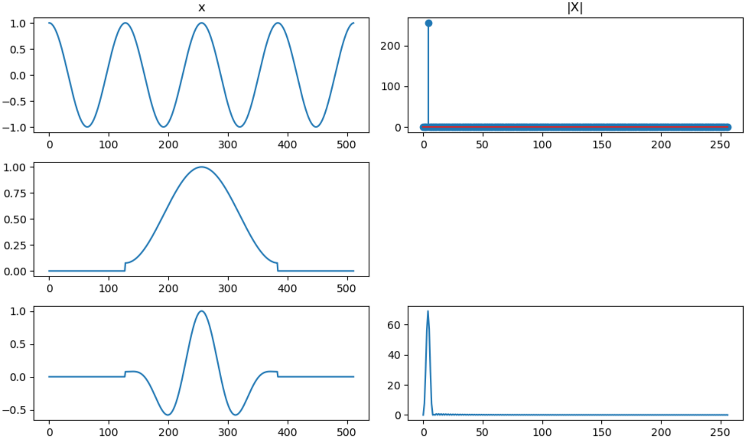


Windowing

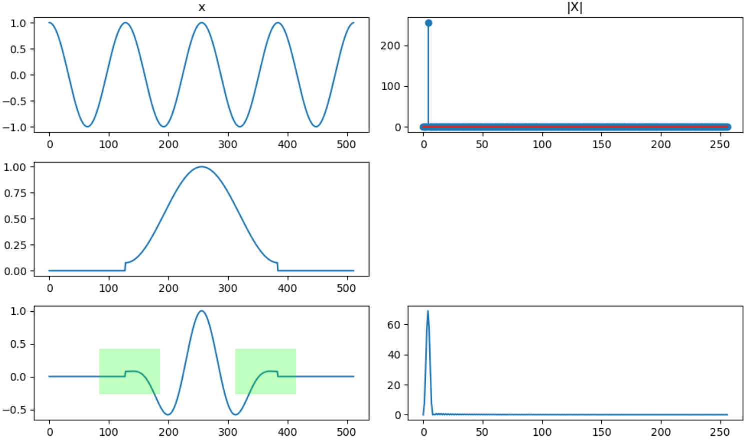


x
|X|
1.0
0.5
0.0
−0.5
−1.0
0
100
200
300
400
500
200
100
0
50
150
250
1.00
0.75
0.50
0.25
0.00

Windowing

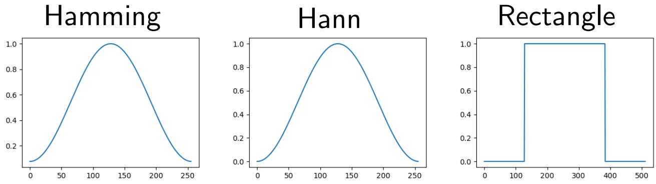

x
|X|
1.0
0.5
0.0
−0.5
−1.0
0
100
200
300
400
500
200
100
0
0
50
100
150
200
250
1.00
0.75
0.50
0.25
0.00
60
40
20

Windowing

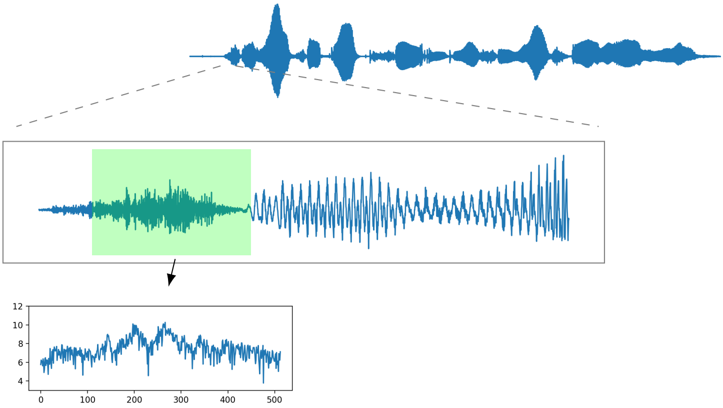

x
|X|
1.0
0.5
0.0
−0.5
−1.0
0
100
200
300
400
500
200
100
0
50
150
250
1.00
0.75
0.50
0.25
0.00
60
40
20

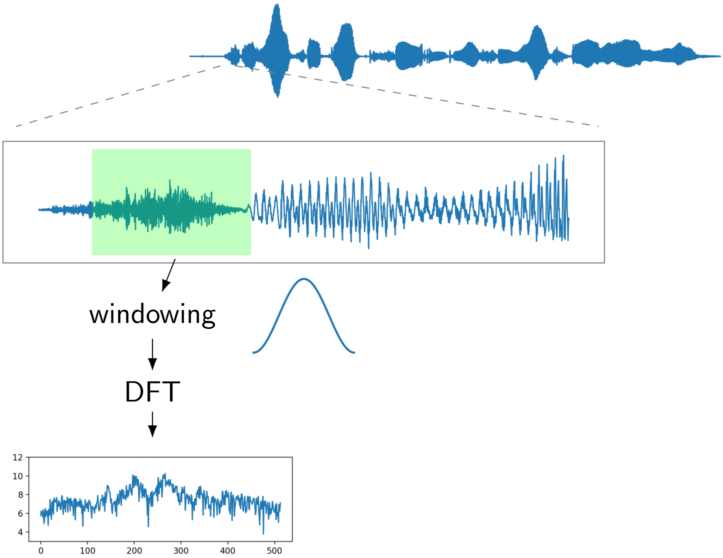
Windowing

$$y[t] = x[t] \cdot w[t]$$

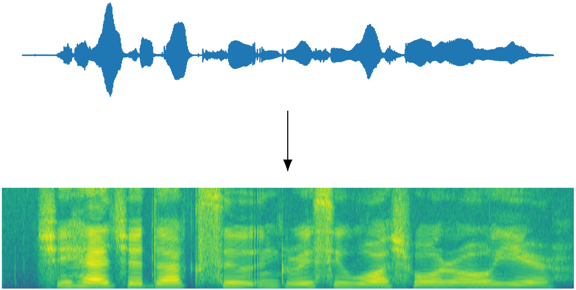


- The signal w is called a window.
- Windowing is elementwise product.


12
10
8
6
4
0
100
200
300
400
500


windowing
DFT
12
10
8
6
4
0
100
200
300
400
500

Spectrogram



- dithering, removing DC offset, pre-emphasis
- windowing
- Discrete Fourier transform (DFT)
- Short-time Fourier transform (STFT)

Further Reading

- Chapter 1–5, Oppenheim, Willsky, and Nawab, "Signals and Systems," 1997
- Chapter 2, O'Shaughnessy, "Speech Communications: Human and Machine," 2000