Speech (Signal) Processing (Part III)

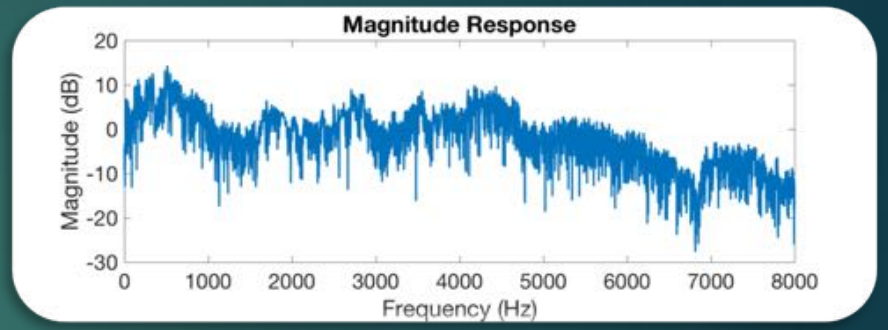
CSCI B659: DEEP LEARNING FOR SPEECH PROCESSING

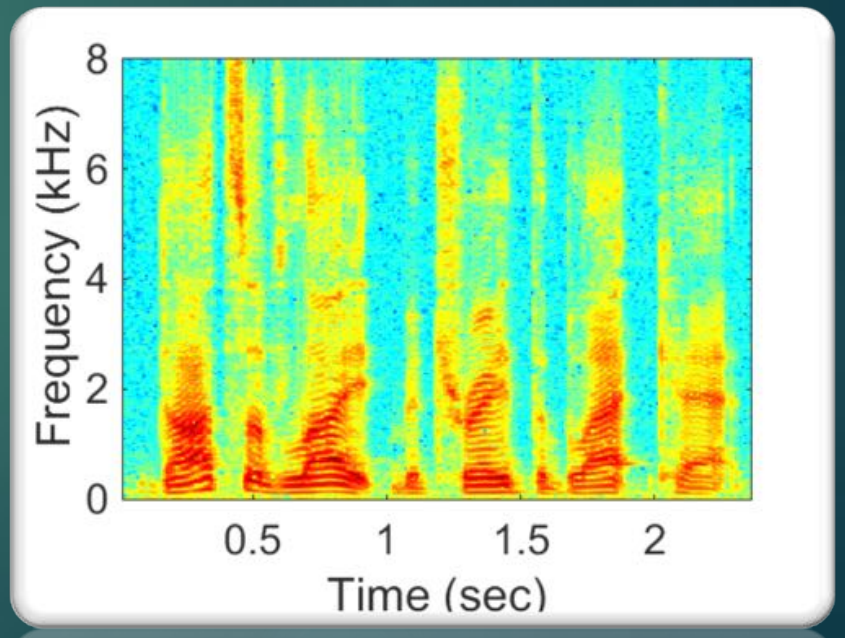
SEPTEMBER 5, 2019

LECTURE #4

Frequency Domain

- Frequency domain on its own may not convey enough information. Why?
  - Frequency information for the entire signal is shown at once
  - Time is still important, but it is lost due to Fourier Analysis
- Time-Frequency (T-F) domain processing is much more widely used





Learning Objectives

- **You the student will be able to:**
  - Explain Time-Frequency Analysis
  - Understand T-F Resolution tradeoffs
  - Explain Time-Frequency Synthesis

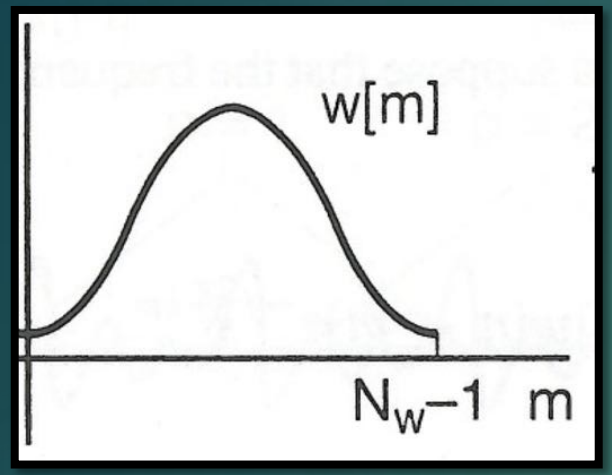
Time-Frequency Domain

- How can you capture information across time and frequency?
  - This is commonly done with the ***Short-time Fourier Transform (STFT)***

$$X[n,k] = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{\frac{-j2\pi km}{N}}$$

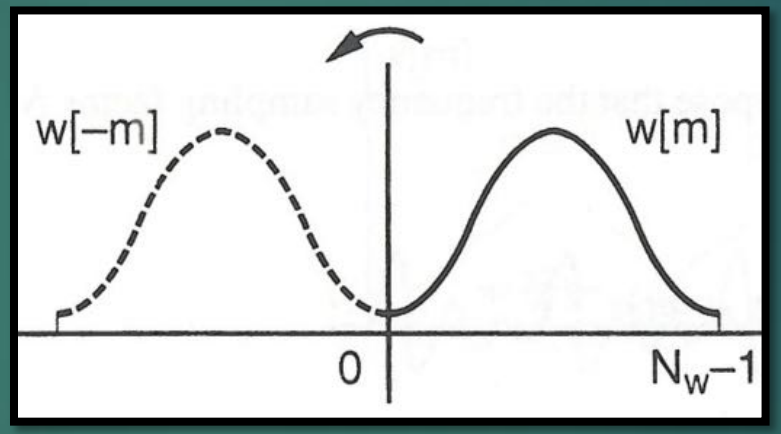
- (Theoretical) Steps:
  1. Define a windowing function w[m]
  2. Flip w[m] along the time axis, w[-m]
  3. Shift w[-m] by n samples, w[n-m]
  4. Multiply w[n-m] and x[m], x[m]w[n-m]
  5. Compute the DFT of x[m]w[n-m]
  6. Repeat 1-5 for all values of n

(Theoretical) STFT Calculation

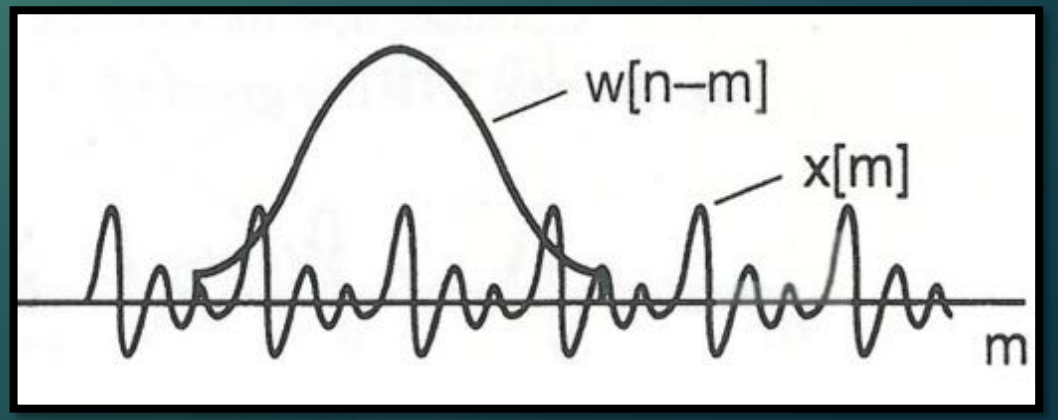


1. Define a windowing function, w[m]

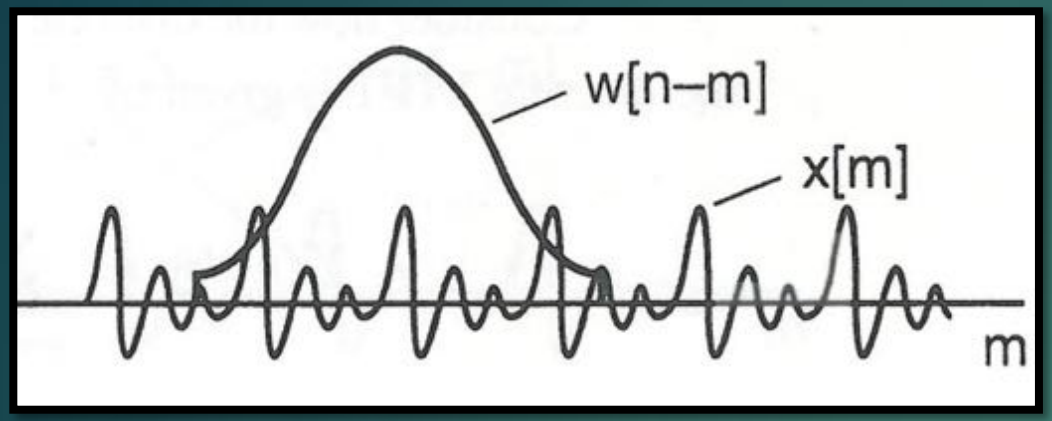
2. Flip w[m] along the time axis, w[-m]



3. Shift w[-m] by n samples, w[n-m]

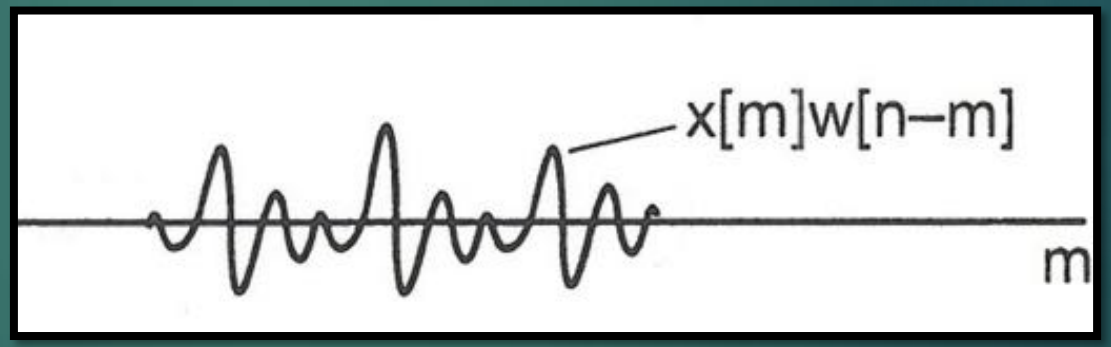


(Theoretical) STFT Calculation (cont.)

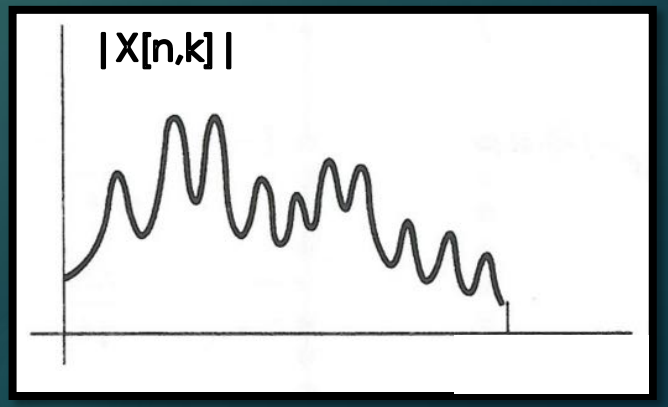


3. Shift w[-m] by n samples, w[n-m]

4. Multiply w[n-m] and x[m]



5. Compute the DFT of x[m]w[n-m]



6. Repeat steps 1-5 for all values of n

Time-Frequency Domain

$$X[n,k] = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{\frac{-j2\pi km}{N}}$$

- **The STFT, *X[n,k]*, is a 2-D matrix**
  - 1st dimension is for **time sample** (i.e. *n*)
  - 2nd dimension is for **frequency index** (i.e. *k*)
  - Each combination of *n* and *k* is referred to as a ***time-frequency (T-F) unit***
- The number of time samples depends on the length of x*[n]*
- The number of frequency channels is N (or *N/2 positive indices)*

Time-Frequency Domain

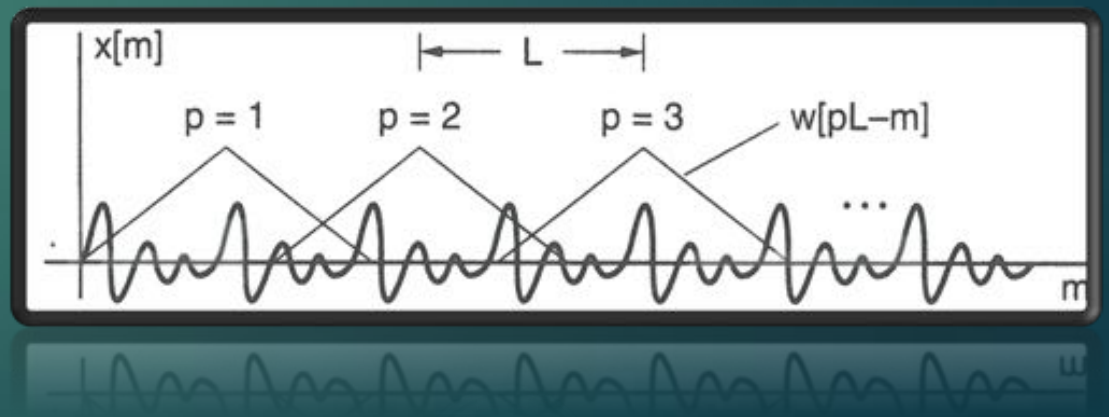
$$X[n,k] = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{\frac{-j2\pi km}{N}}$$

- Typically, the STFT is ***NOT*** computed at all time samples *n*
- Usually, the windowing function is shifted by *L* samples, and not by *1*
  - *L* is the window shift amount
  - Shifting by *1* sample creates unnecessary data

Time-Frequency Domain

$$X[p,k] = \sum_{m=-\infty}^{\infty} x[m]w[pL-m]e^{\frac{-j2\pi km}{N}}$$

- Steps:
  1. Define a windowing function *w[m]*
  2. Flip *w[m]* along the time axis, *w[-m]*
  3. **Shift *w[-m]* by *pL* samples, *w[pL-m]***
  4. Multiply *w[pL-m]* and *x[m]*, *x[m]w[pL-m]*
  5. Compute the DFT of *x[m]w[pL-m]*
  6. **Repeat 1-5 for all values of *p***



STFT Example

- See Demo

STFT Calculation in Practice

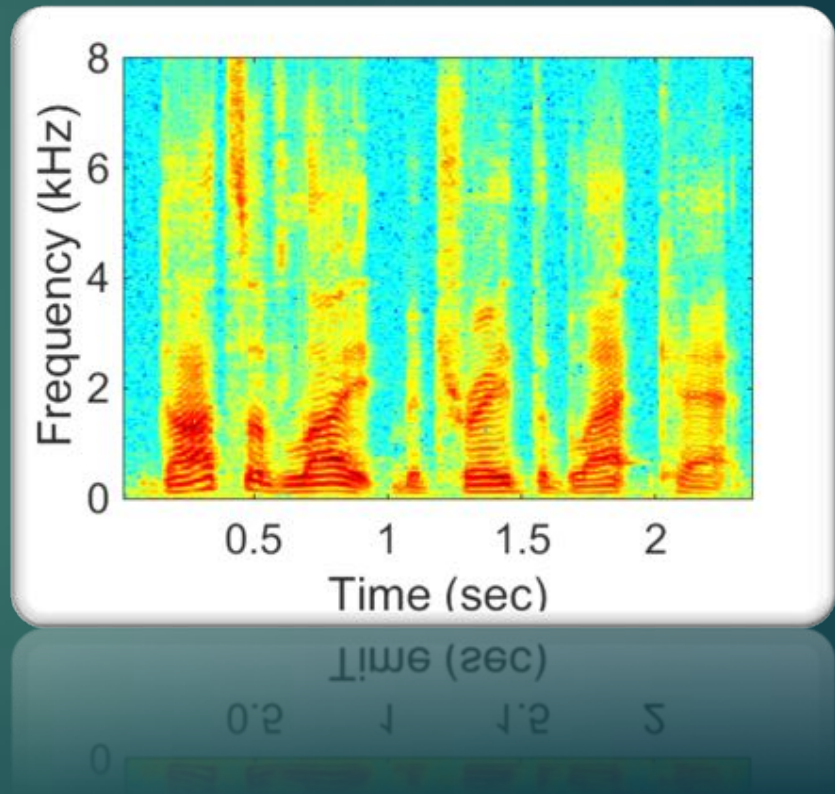
$$X[p,k] = \sum_{m=-\infty}^{\infty} x[m]w[pL-m]e^{\frac{-j2\pi km}{N}}$$

- How is the STFT calculated in practice?
  - Offline version (Have access to full audio signal)
  - Online version (Have access to portion of audio signal)

- See board

Spectrogram

- Plotting the $|X[p,k]|$ for all p and k, in a single plot, results in a ***spectrogram***
- This is often plotted in dB. Hence, $10\log_{10}(|X[p,k]|)$
- The ***power spectrogram*** results when plotting $20\log_{10}(|X[p,k]|)$
  - Why is this name used?



Recap

- The time-frequency (T-F) domain provides details of how a signal changes over time and frequency
  - This information is available in the other domains
  - The need for this will come apparent later (speech production and recognition)

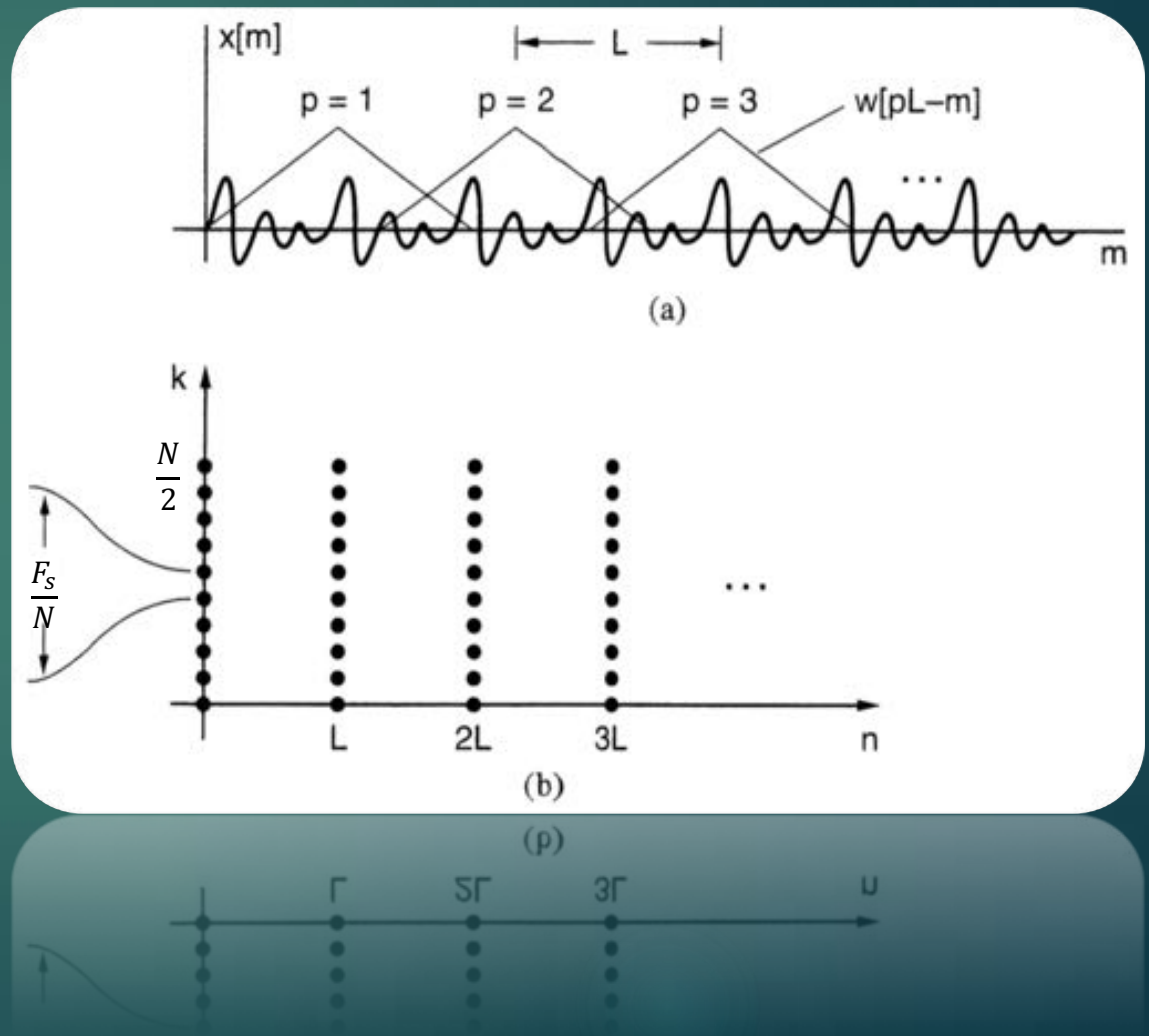
- The ***short-time Fourier transform*** (STFT) is the method to compute the T-F signal
  - Fourier Transform view
  - Filtering view

Learning Objectives

- **You the student will be able to:**
  - Explain Time-Frequency Analysis
  - Understand T-F Resolution tradeoffs
  - Explain Time-Frequency Synthesis

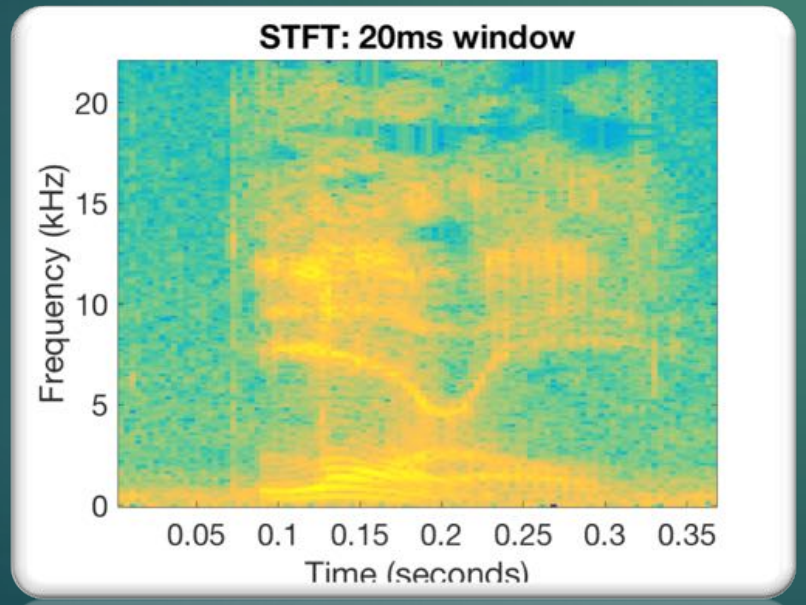
Time-Frequency Sampling

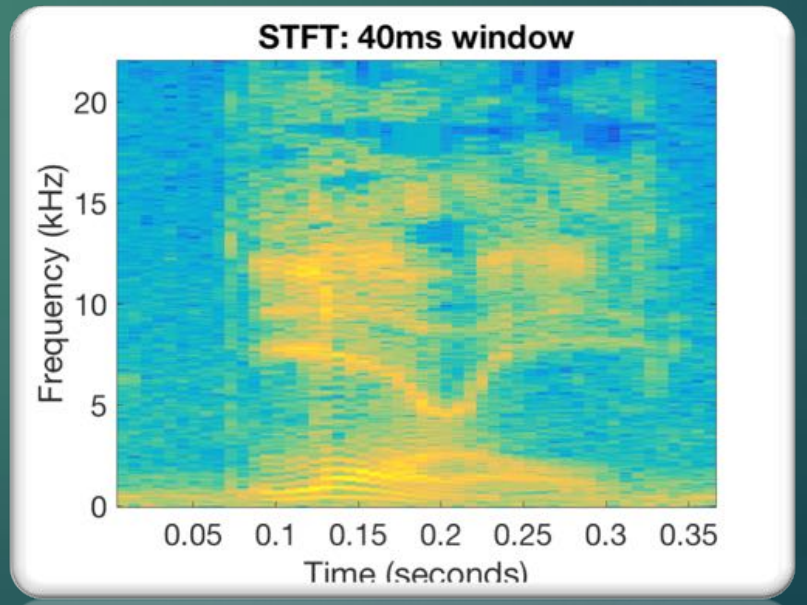
- The STFT does not give values at all time and frequency points
- The time values depend on the window hopsize (*L*) and *p*
- The frequency values depend on *N* and F_s (sampling rate)



T-F Resolution Tradeoff

- The length of the analysis window affects T-F resolution
  - ***Long window length***: Fine frequency detail, poor temporal structure
  - ***Short window length***: Poor frequency detail, fine temporal structure
- More on this later





Learning Objectives

- **You the student will be able to:**
  - Explain Time-Frequency Analysis
  - Understand T-F Resolution tradeoffs
  - Explain Time-Frequency Synthesis

Synthesis: T-F to Time domain

Overlap-Add (OLA)

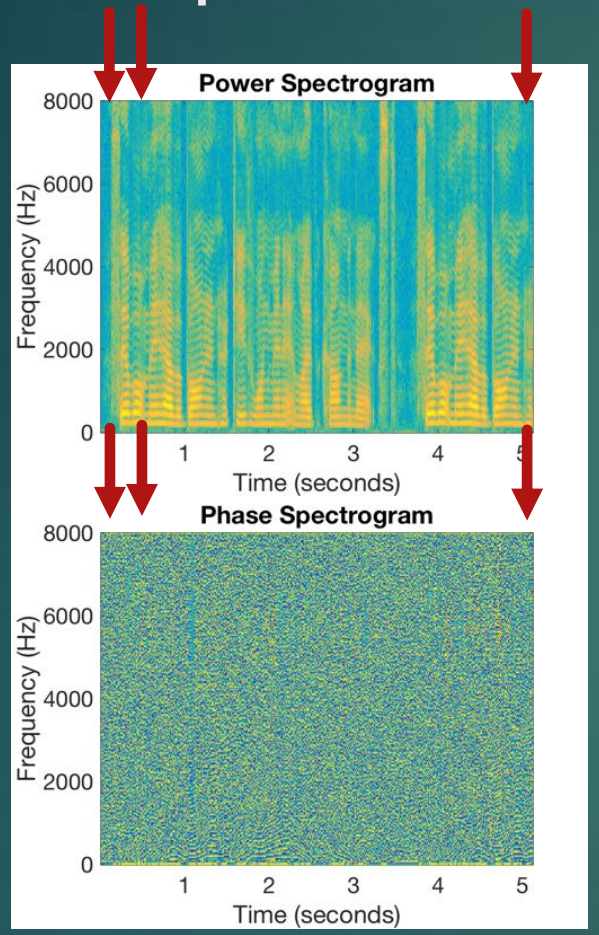
$$x[n] = \frac{L}{W[0]} \sum_{p=-\infty}^{\infty} \left[\frac{1}{N} \sum_{k=0}^{N-1} X[pL, k] e^{j \frac{2\pi}{N} kn} \right]$$

$x[n]w[pL - n]$

- Fourier transform approach
- Adding time components for each *n*
- Steps:
  - Take inverse DFT for each time frame
  - Sum over all *p*
  - Scale by a factor
  - Repeat for all *n*

Overlap-Add Synthesis

$$x[n] = \frac{L}{W[0]} \sum_{p=-\infty}^{\infty} \left[\frac{1}{N} \sum_{k=0}^{N-1} X[pL, k] e^{j \frac{2\pi}{N} kn} \right]$$

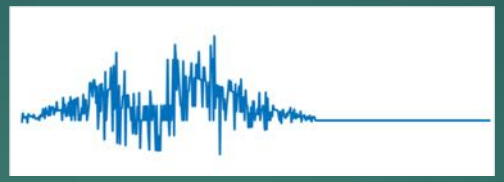


1. Combine Magnitude and Phase Spectrograms to STFT

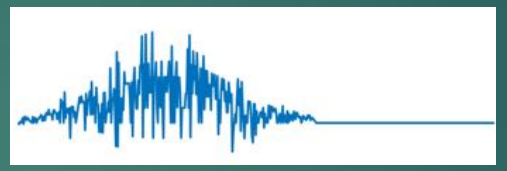
=

0

N-1



+

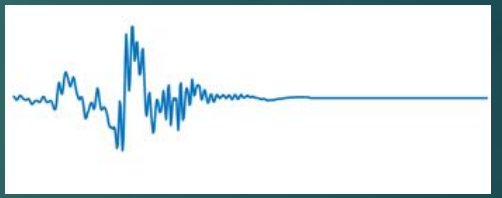


L

L+N - 1

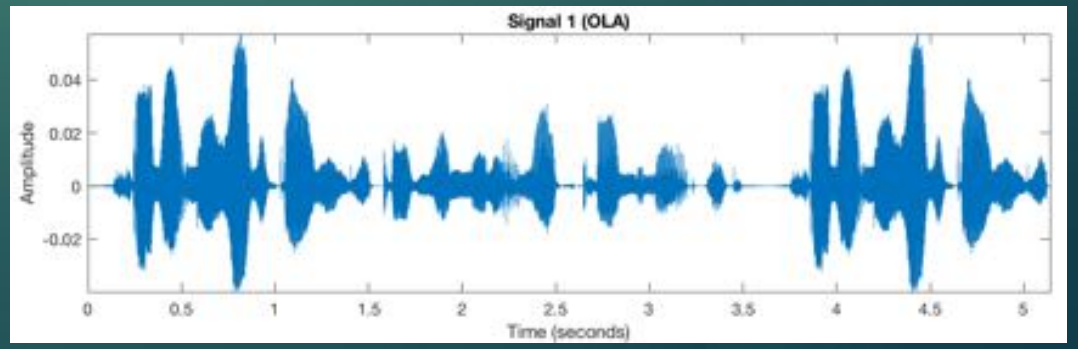
⋱

+



2. Compute IDFT of each frame (e.g. time frame) and place at appropriate sample position

3. Add across all frames



- Now: In-class (Homework) Assignment
- Next Class
  - Signal Processing (Part IV)
    - Filtering
    - Analysis Windows