

# Speech Technology: A Practical Introduction

## Topic: Spectrogram, Cepstrum and Mel-Frequency Analysis

Kishore Prahallad  
Email: [skishore@cs.cmu.edu](mailto:skishore@cs.cmu.edu)

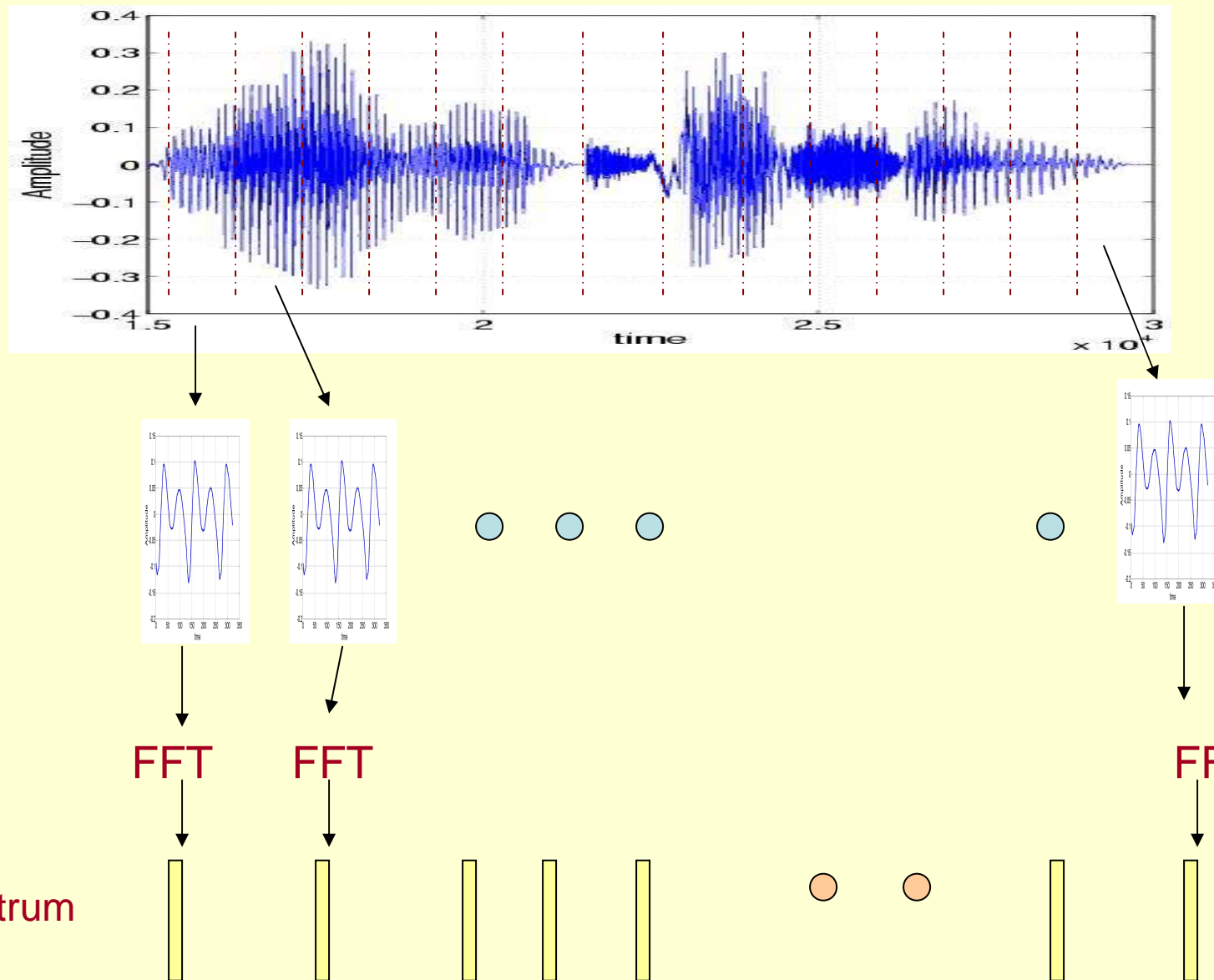
Carnegie Mellon University  
&  
International Institute of Information Technology Hyderabad

# Topics

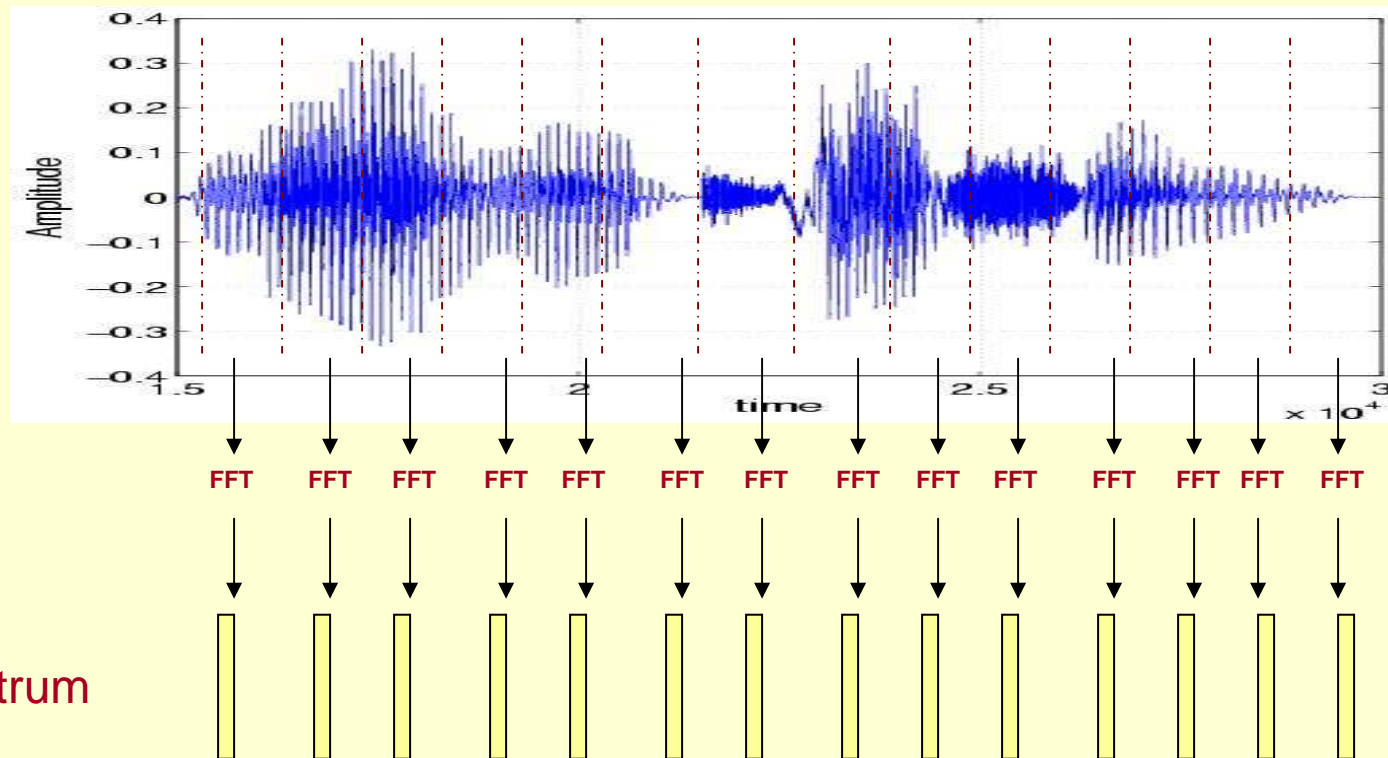
- Spectrogram
- Cepstrum
- Mel-Frequency Analysis
- Mel-Frequency Cepstral Coefficients

# Spectrogram

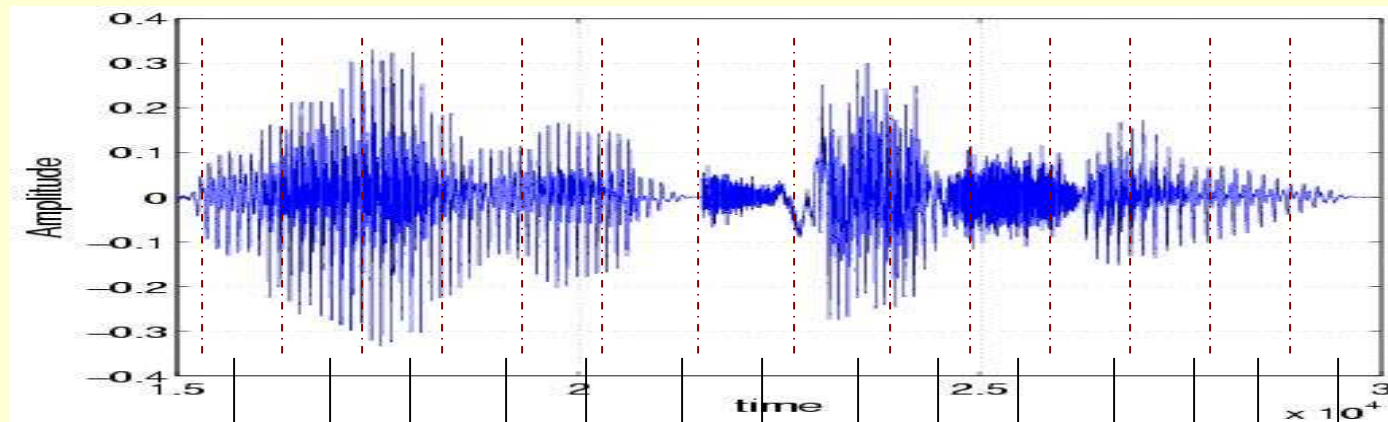
# Speech signal represented as a sequence of spectral vectors



# Speech signal represented as a sequence of spectral vectors

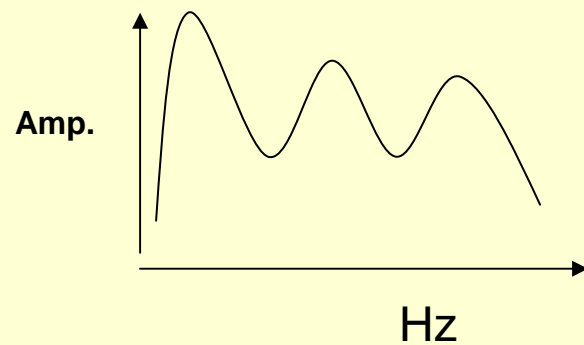
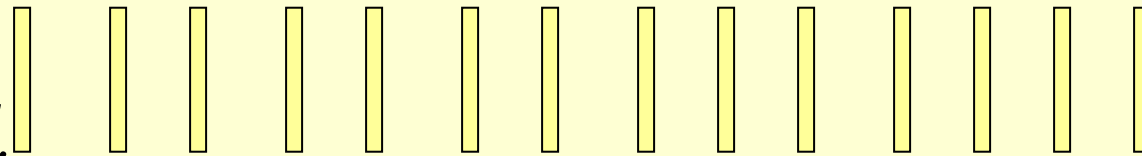


# Speech signal represented as a sequence of spectral vectors

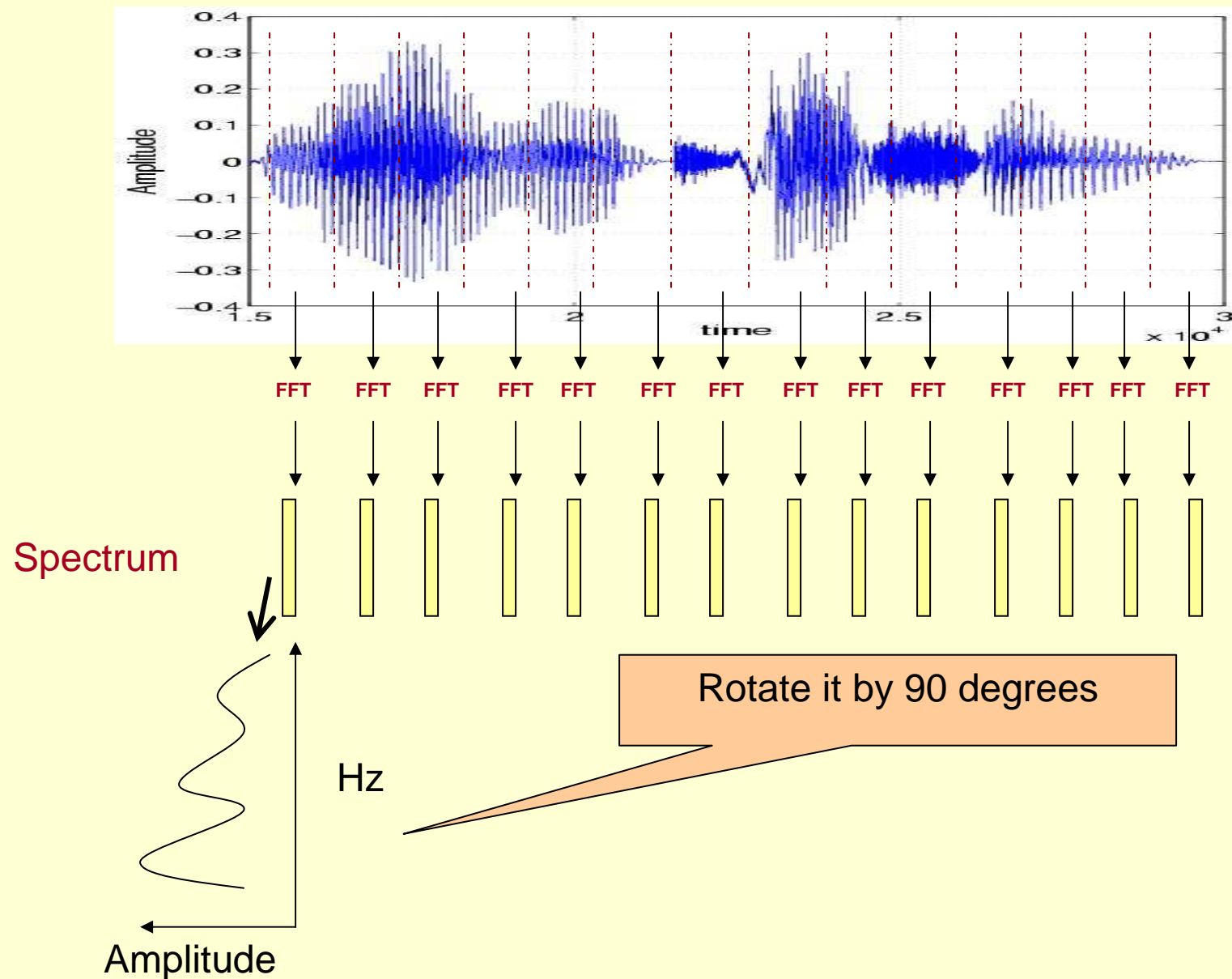


FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT

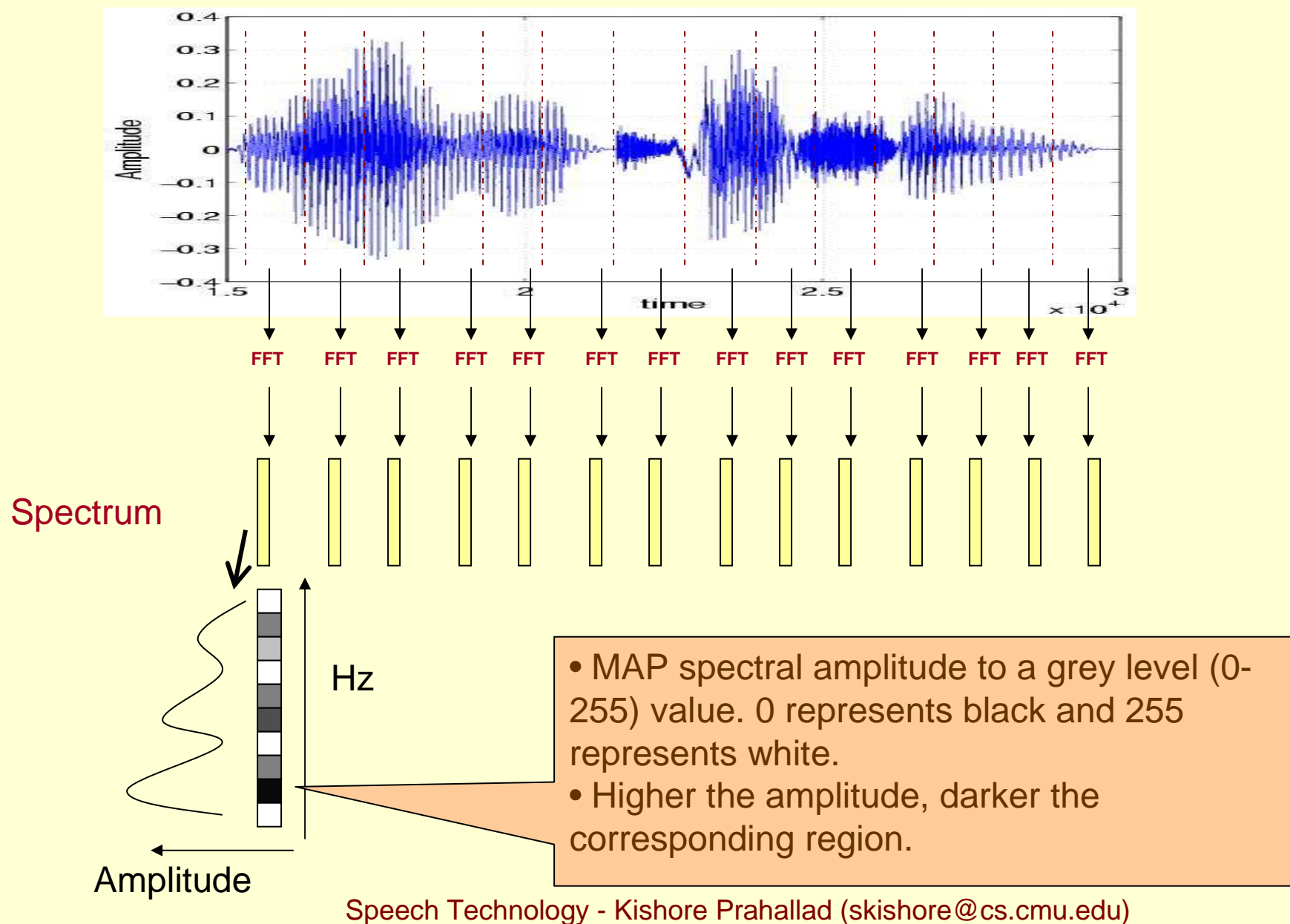
Spectrum



# Speech signal represented as a sequence of spectral vectors

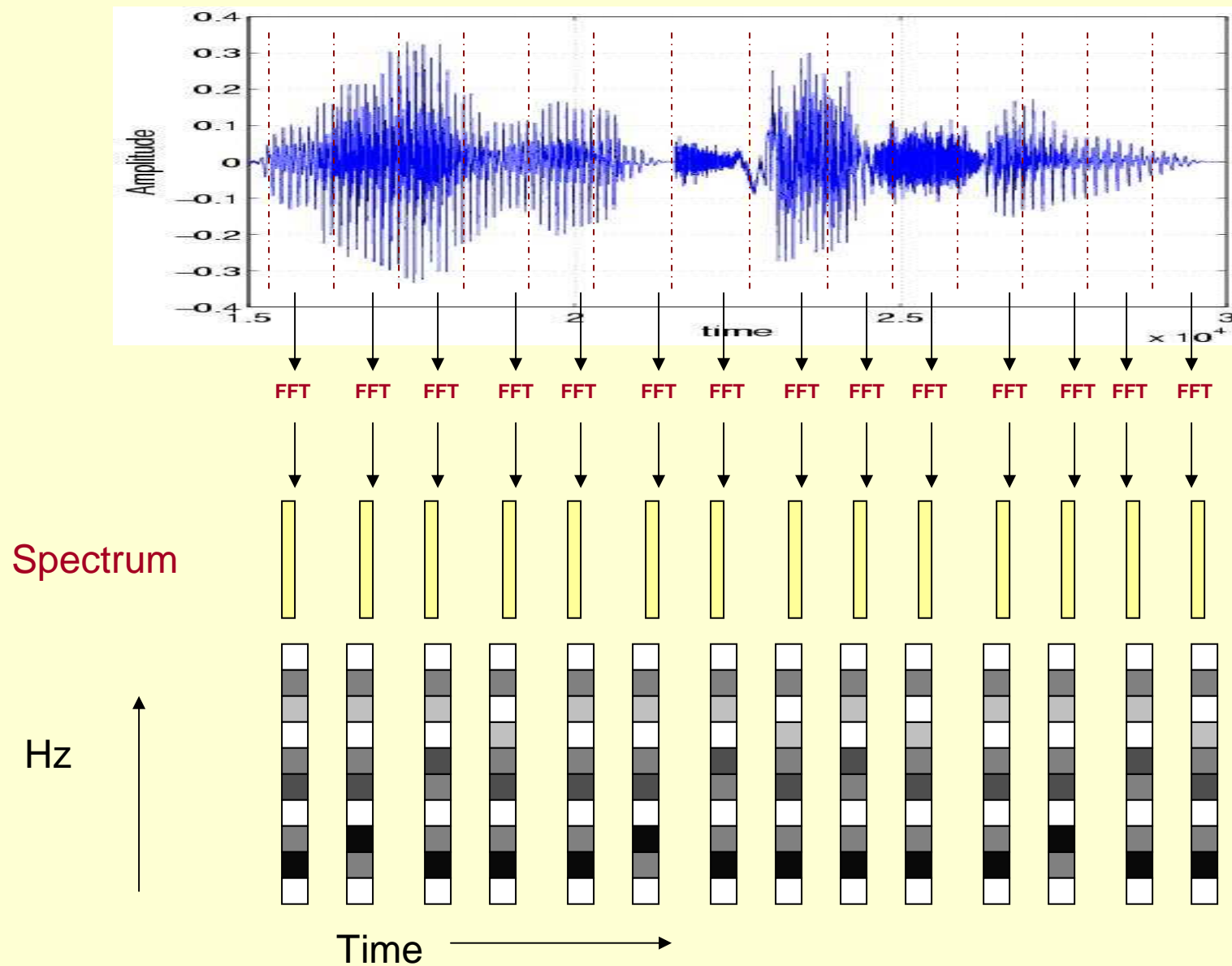


# Speech signal represented as a sequence of spectral vectors

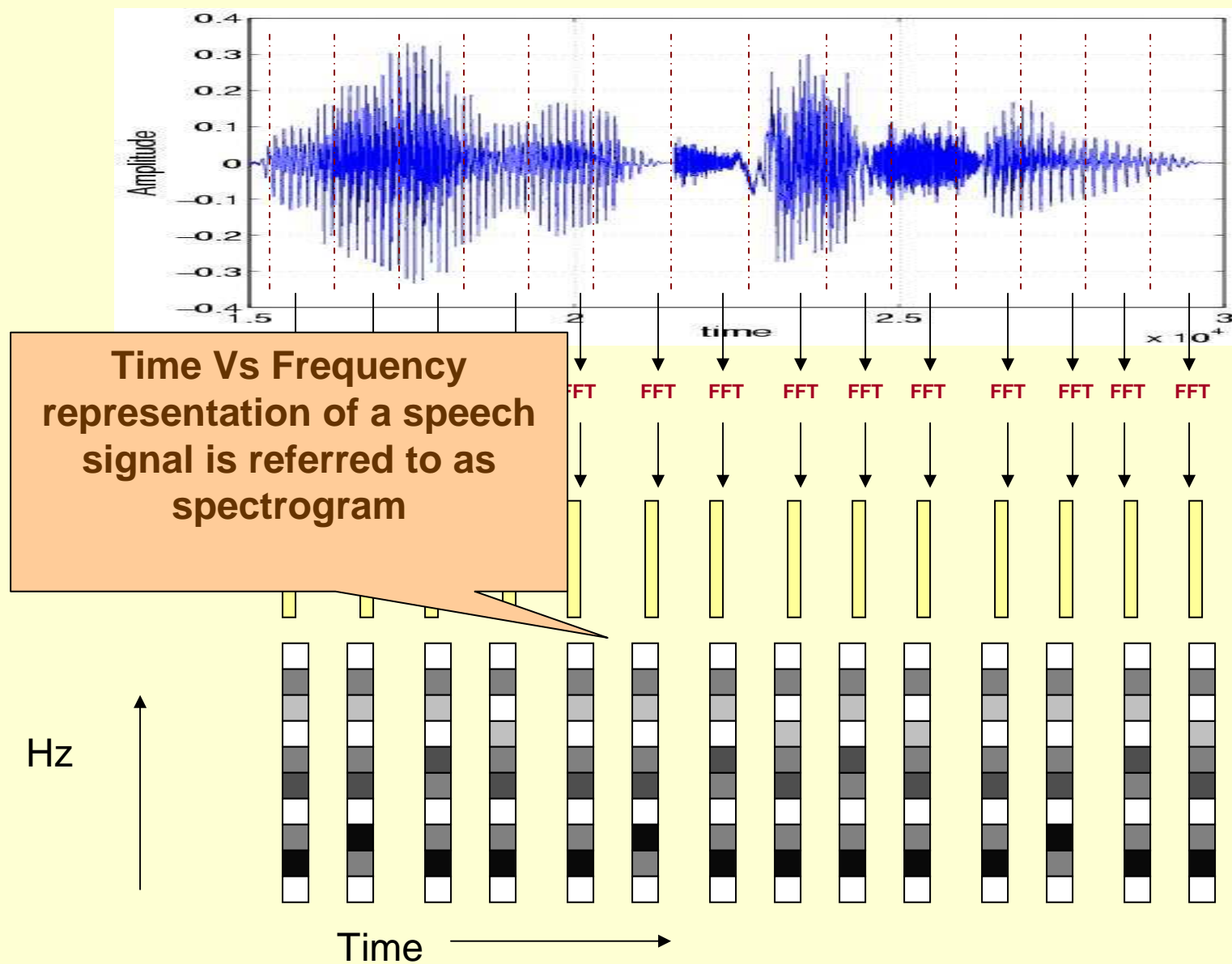




# Speech signal represented as a sequence of spectral vectors

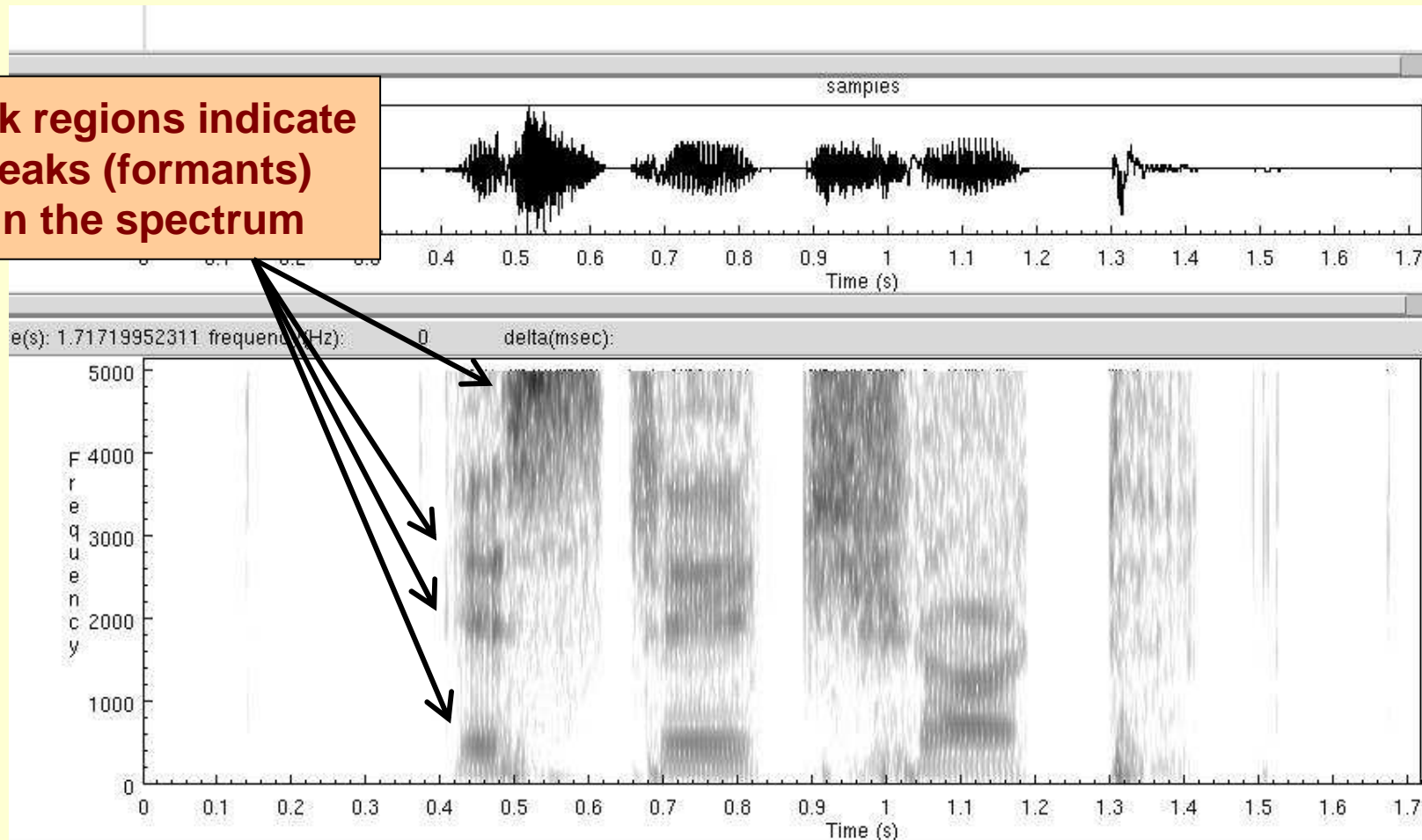


# Speech signal represented as a sequence of spectral vectors



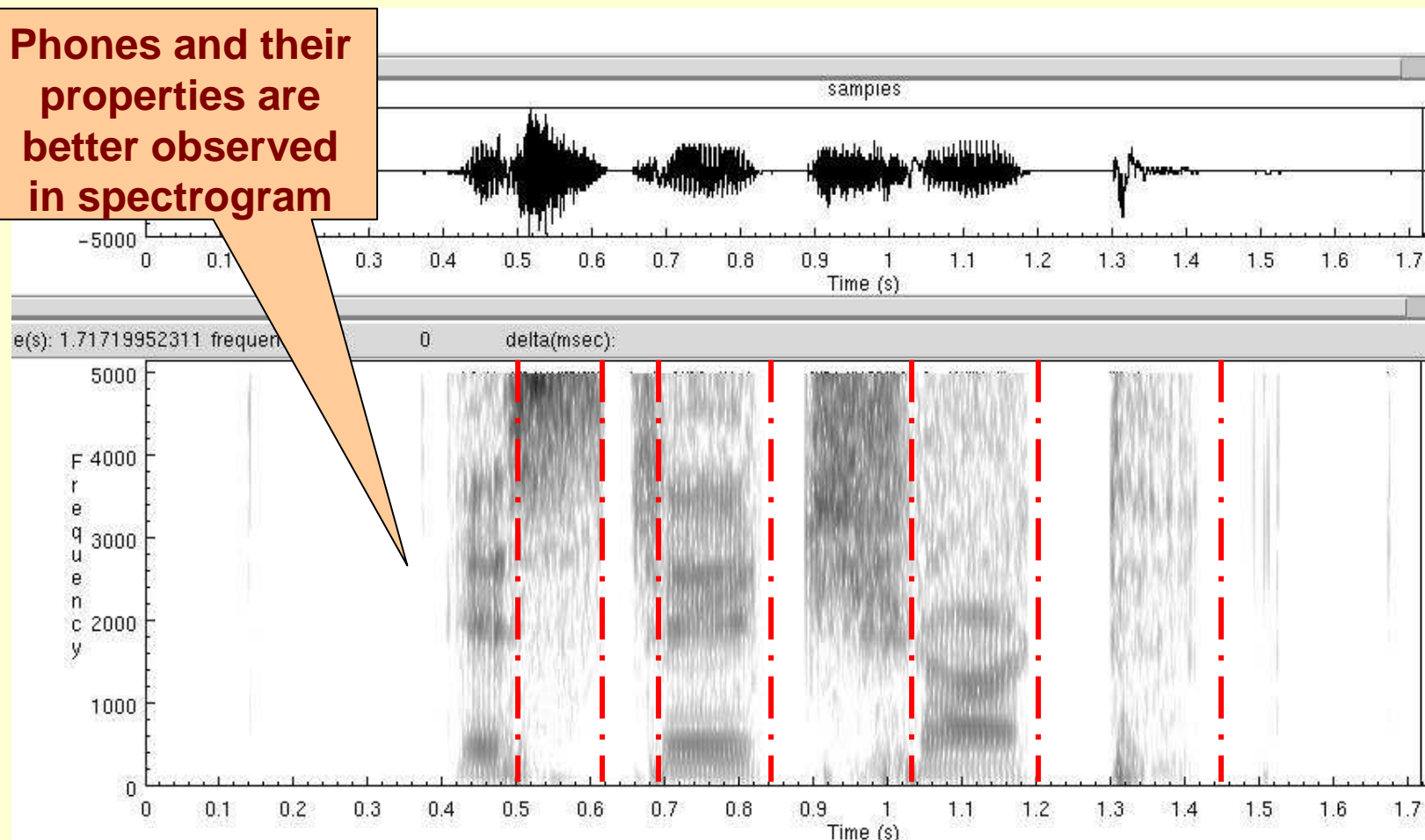
# Some Real Spectrograms

**Dark regions indicate  
peaks (formants)  
in the spectrum**



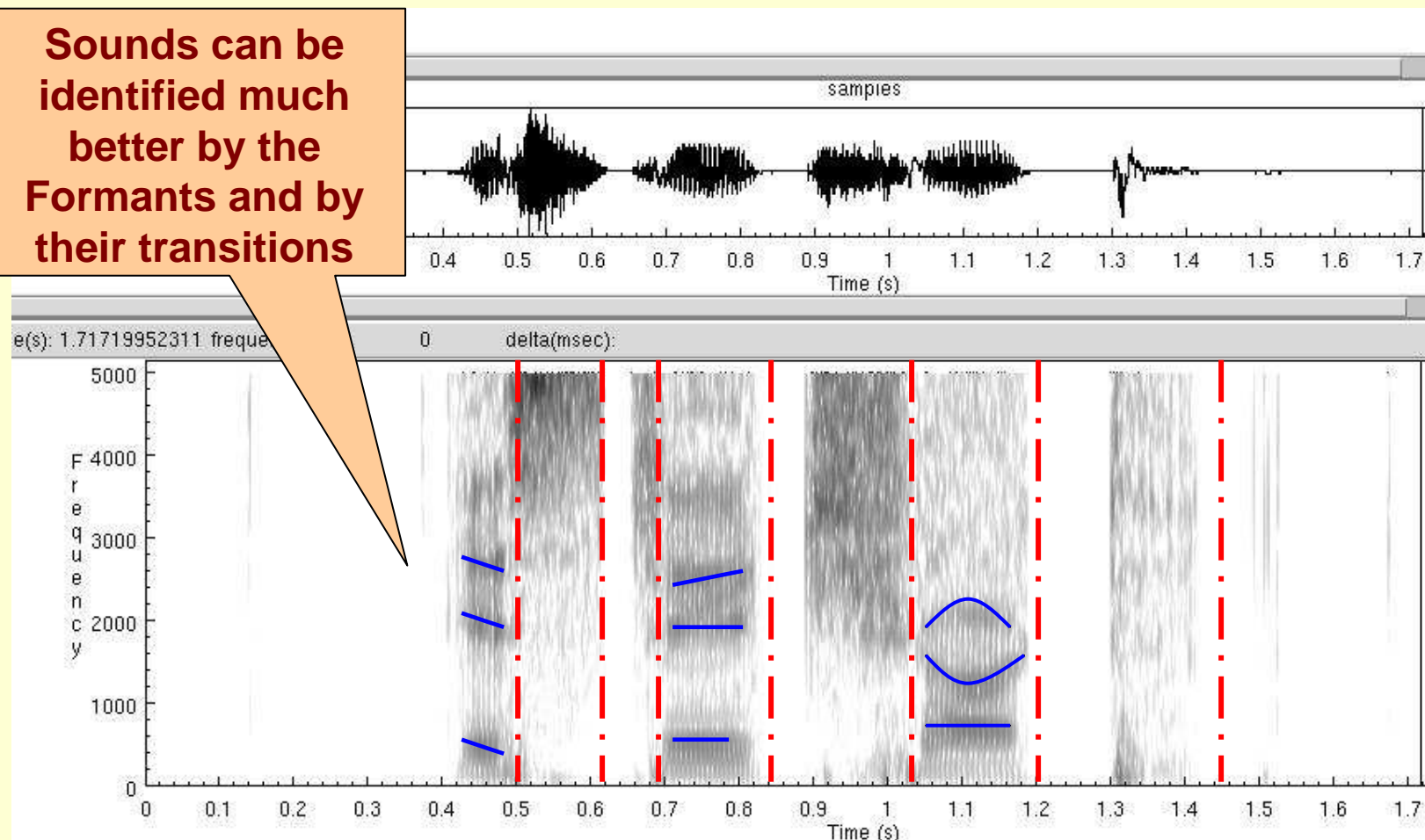
# Why we are bothered about spectrograms

Phones and their properties are better observed in spectrogram



# Why we are bothered about spectrograms

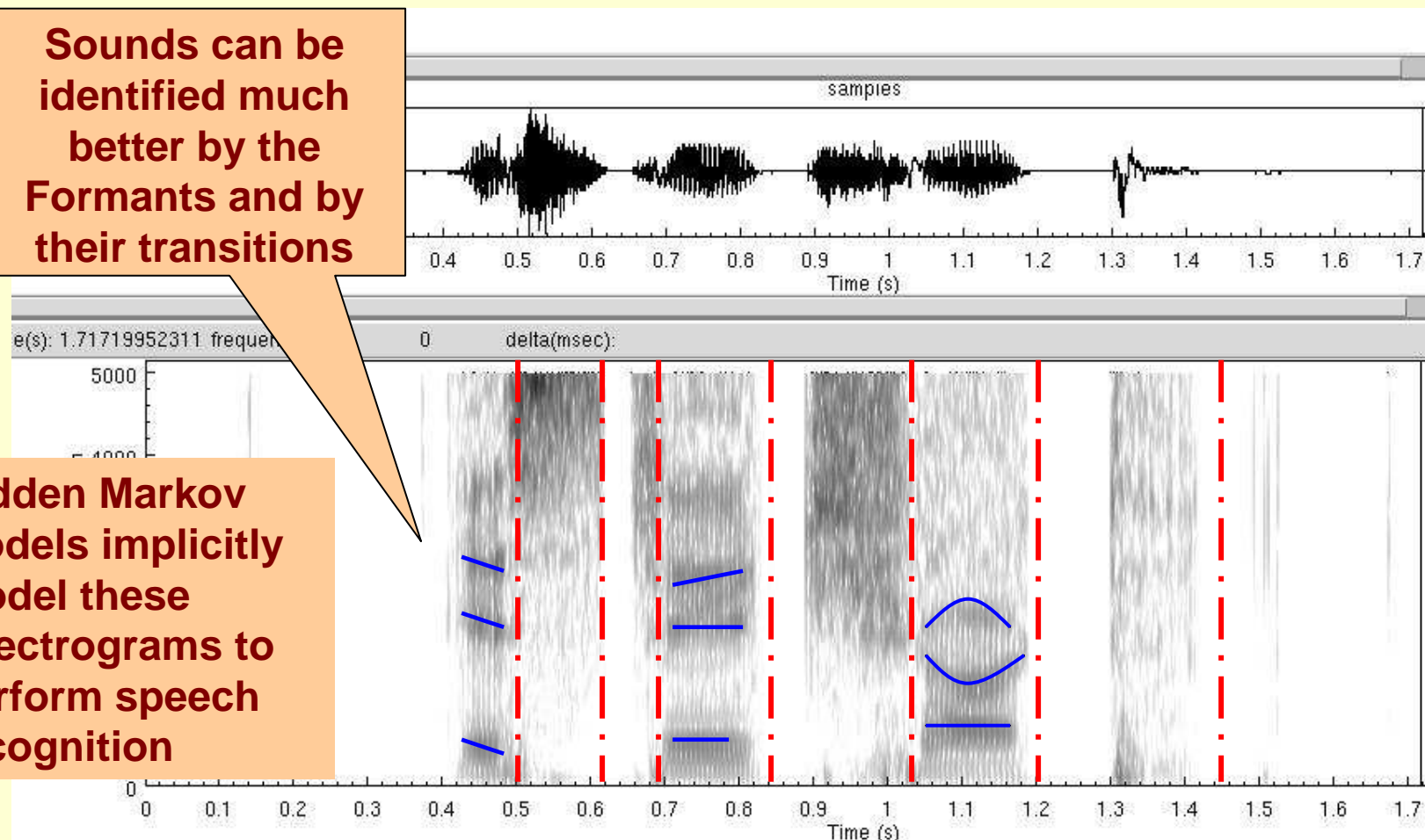
Sounds can be identified much better by the Formants and by their transitions



# Why we are bothered about spectrograms

Sounds can be identified much better by the Formants and by their transitions

Hidden Markov Models implicitly model these spectrograms to perform speech recognition



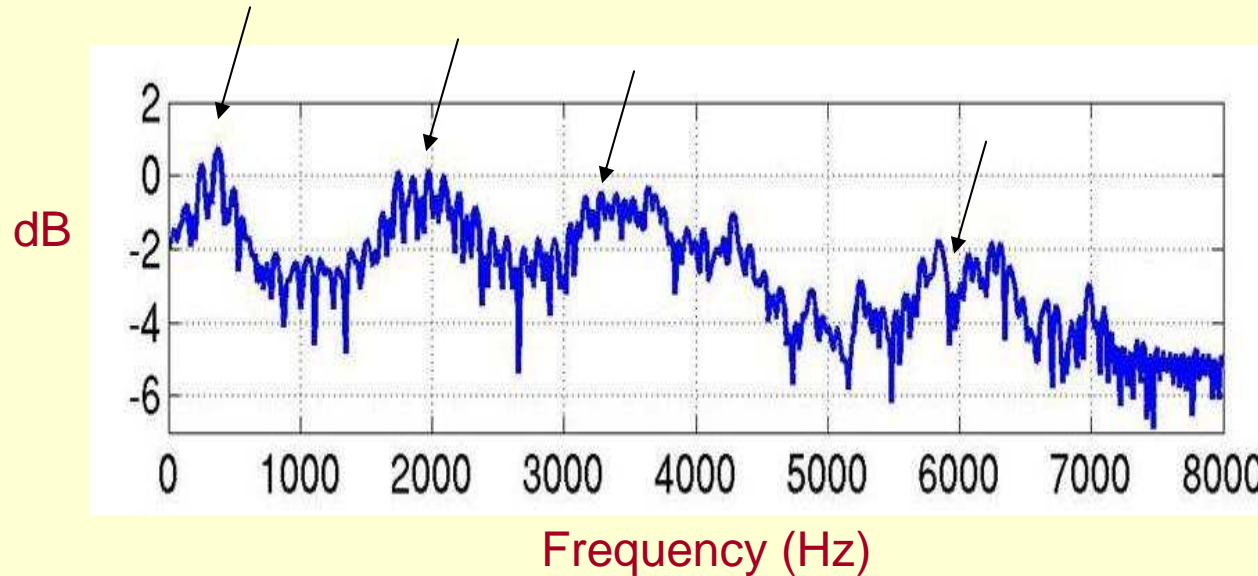
# Usefulness of Spectrogram

- Time-Frequency representation of the speech signal
- Spectrogram is a tool to study speech sounds (phones)
- Phones and their properties are visually studied by phoneticians
- Hidden Markov Models implicitly model spectrograms for speech to text systems
- Useful for evaluation of text to speech systems
  - A high quality text to speech system should produce synthesized speech whose spectrograms should nearly match with the natural sentences.

# Cepstral Analysis



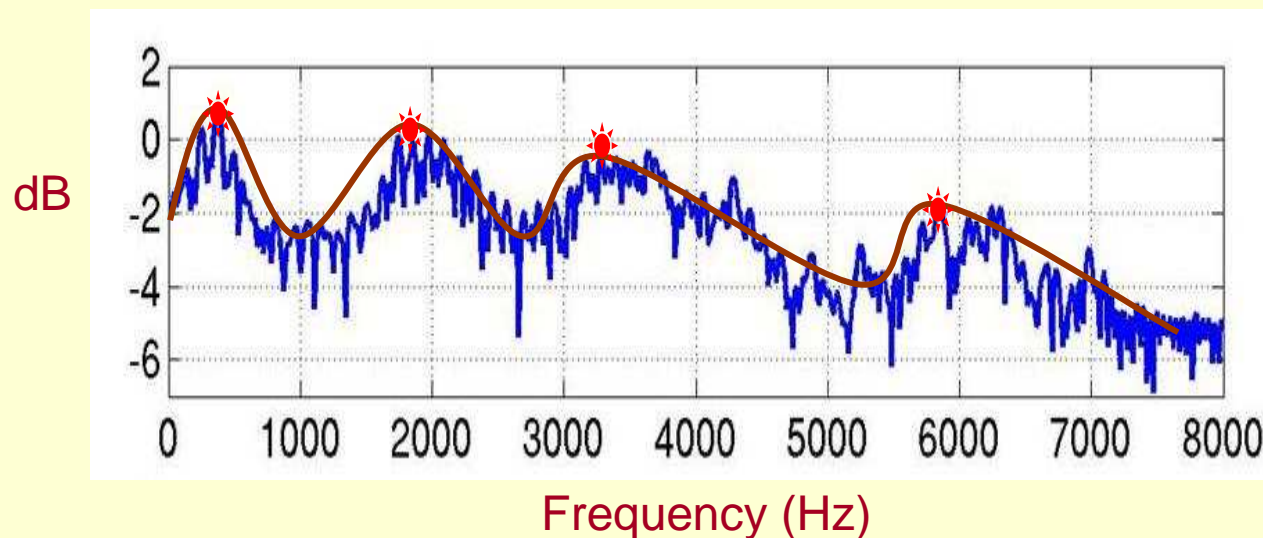
# A Sample Speech Spectrum



- Peaks denote dominant frequency components in the speech signal
- Peaks are referred to as formants
- Formants carry the identity of the sound

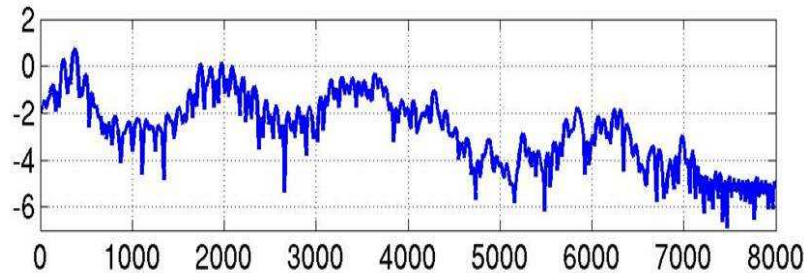
# What we want to Extract? – Spectral Envelope

- Formants and a smooth curve connecting them
- This Smooth curve is referred to as spectral envelope

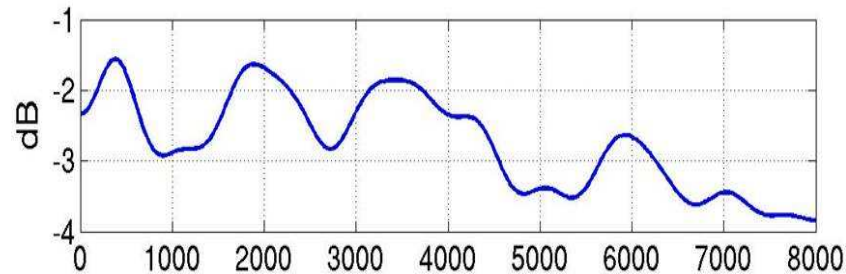


# Spectral Envelope

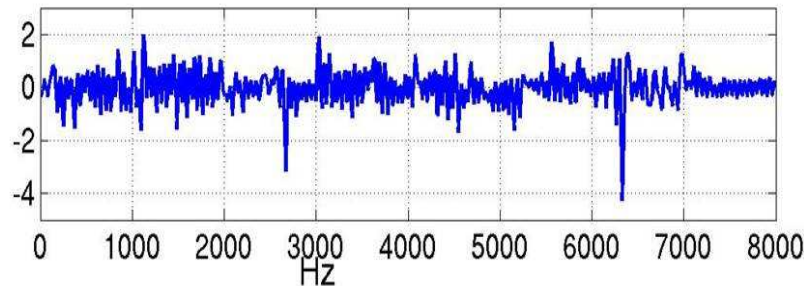
**Spectrum**



**Spectral  
Envelope**

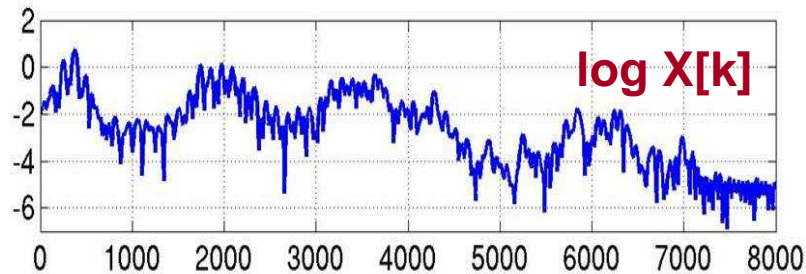


**Spectral  
details**

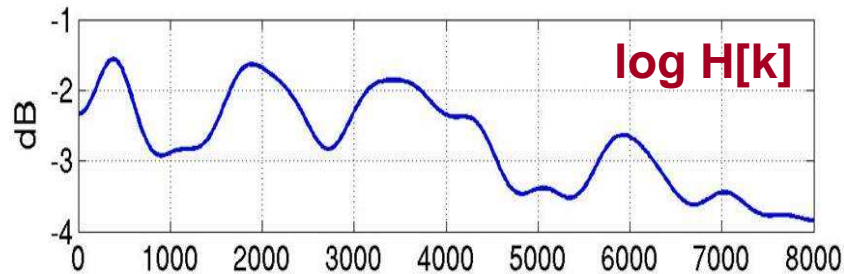


# Spectral Envelope

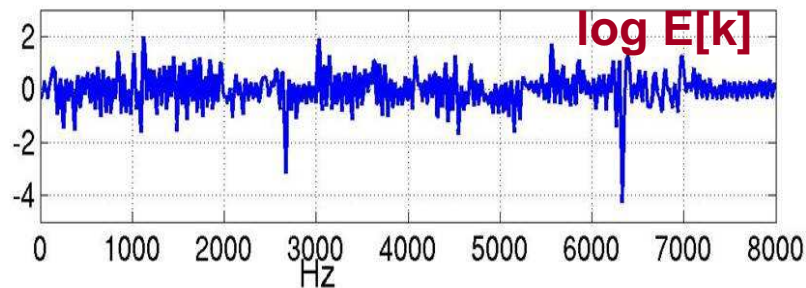
**Spectrum**



**Spectral Envelope**

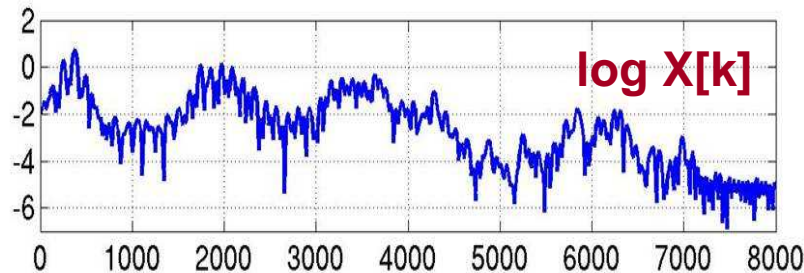


**Spectral details**

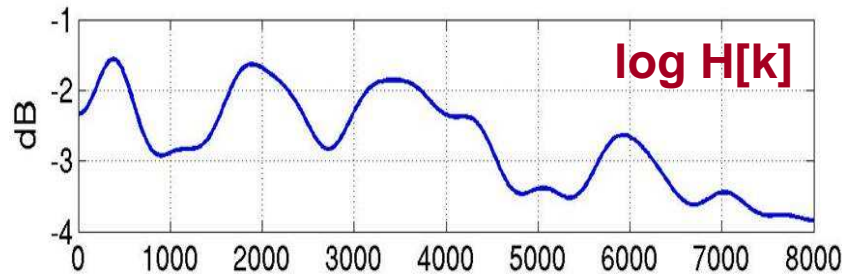


# Spectral Envelope

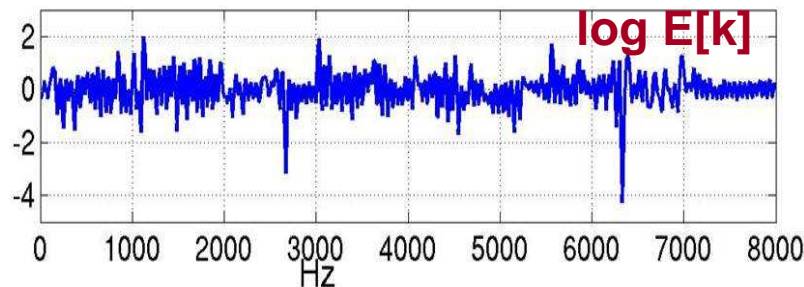
**Spectrum**



**Spectral Envelope**



**Spectral details**



$$\log X[k] = \log H[k] + \log E[k]$$

1. Our goal: We want to separate spectral envelope and spectral details from the spectrum.

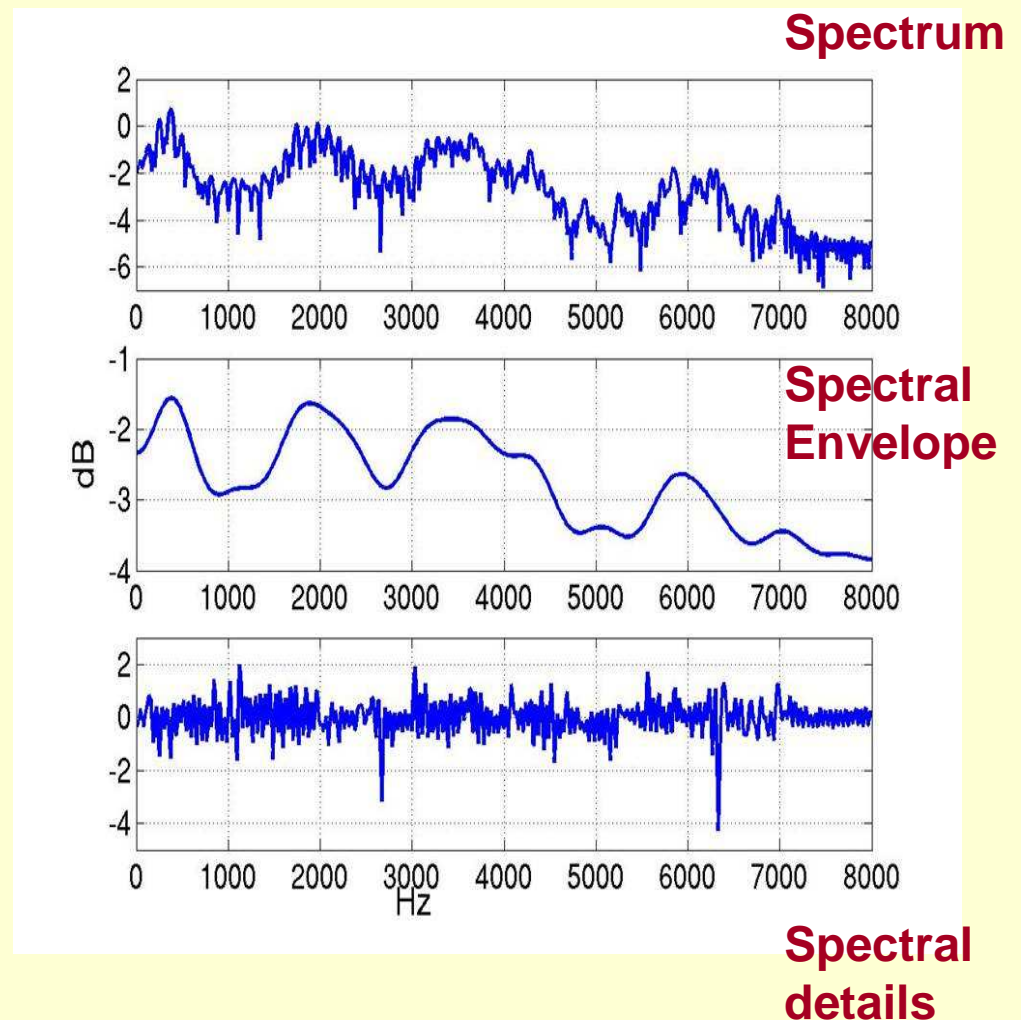
2. i.e Given  $\log X[k]$ , obtain  $\log H[k]$  and  $\log E[k]$ , such that  $\log X[k] = \log H[k] + \log E[k]$

How to achieve this  
separation ?



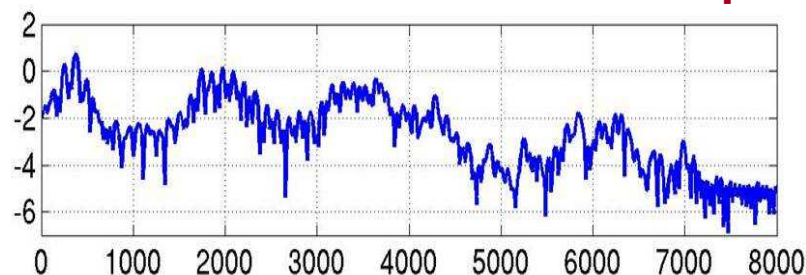
# Play a Mathematical Trick

- Trick: Take FFT of the spectrum!!
- An FFT on spectrum referred to as Inverse FFT (IFFT).
- Note: We are dealing with spectrum in log domain (part of the trick)
- IFFT of log spectrum would represent the signal in pseudo-frequency axis

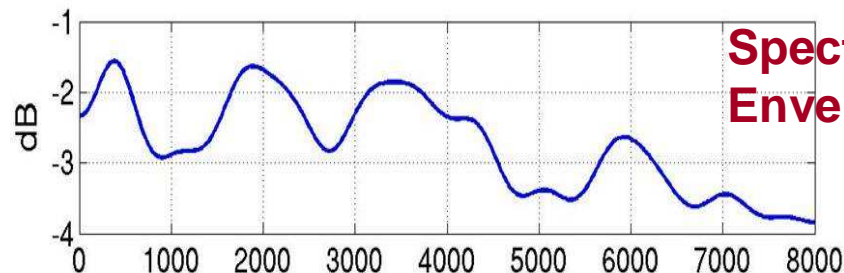


# Play a Mathematical Trick

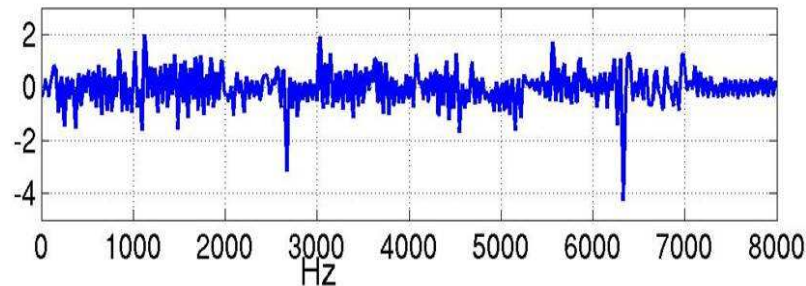
**Spectrum**



**Spectral Envelope**



**Spectral details**



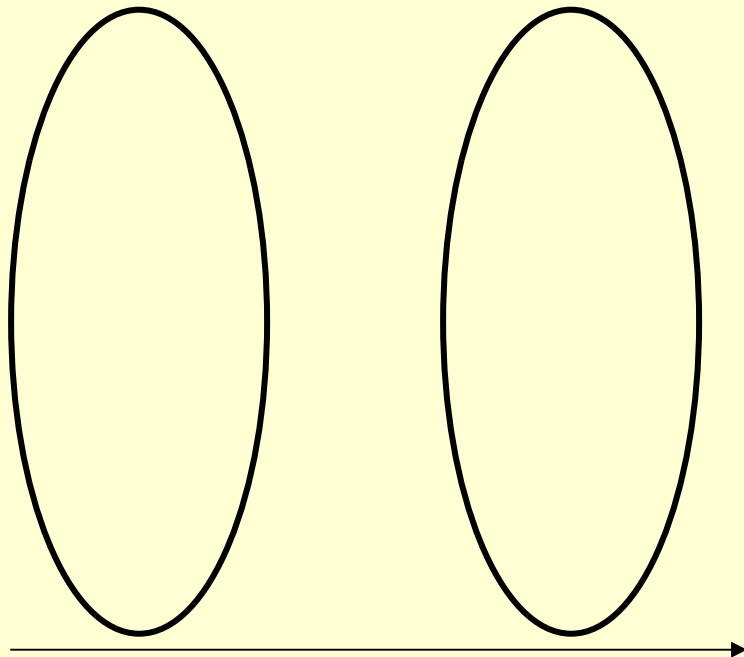
**A pseudo-frequency axis**



# Play a Mathematical Trick

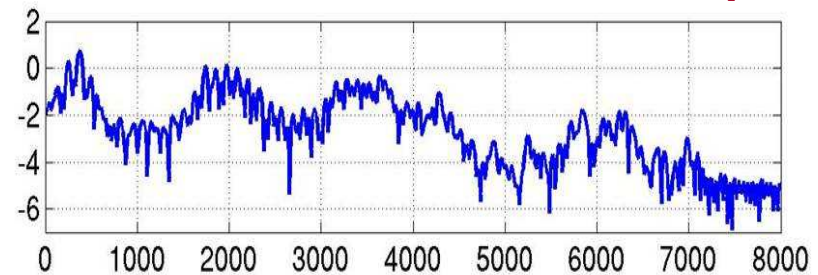
Low Freq.  
region

High Freq.  
region

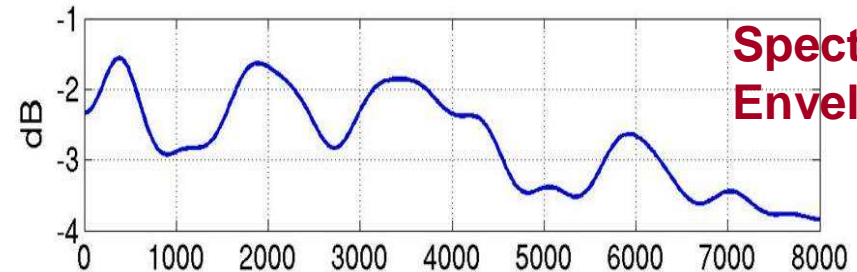


A pseudo-frequency  
axis

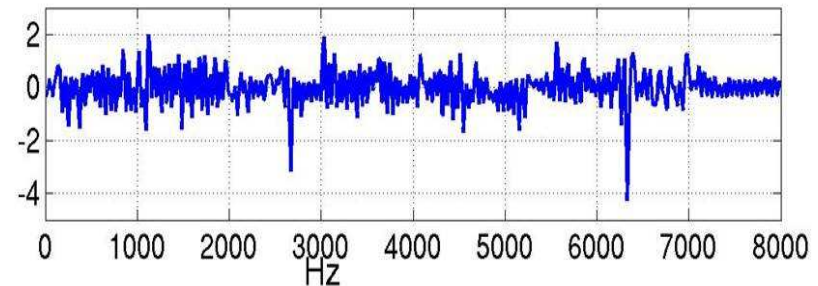
Spectrum



Spectral  
Envelope



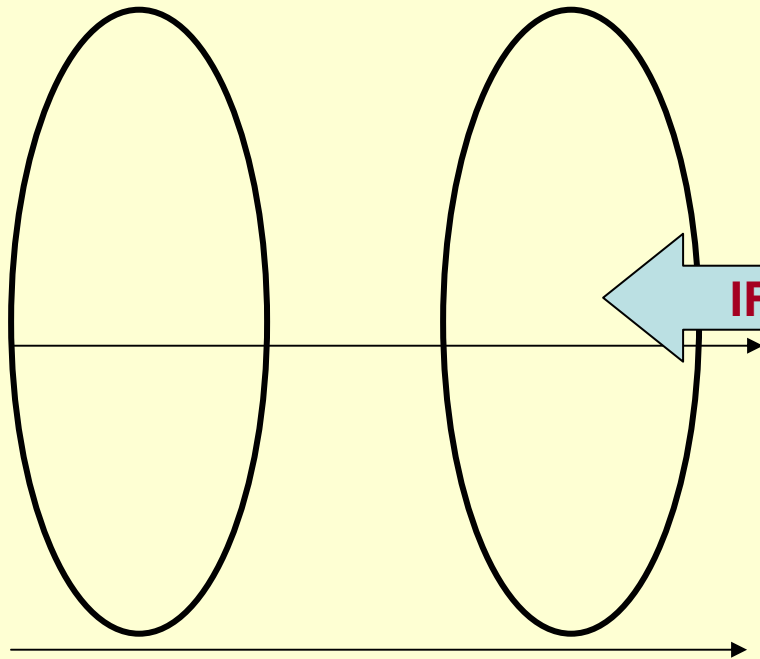
Spectral  
details



# Play a Mathematical Trick

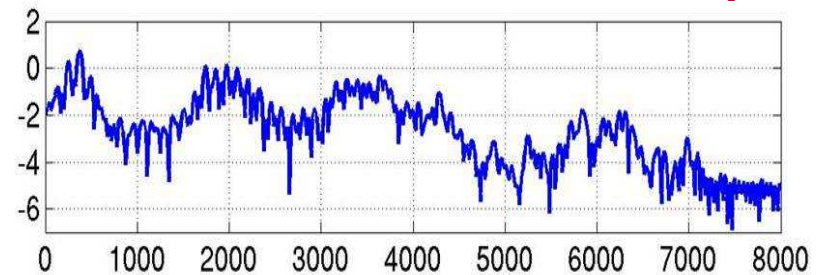
Low Freq.  
region

High Freq.  
region

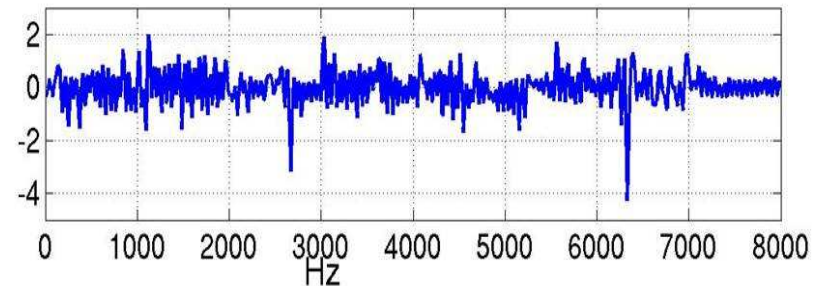
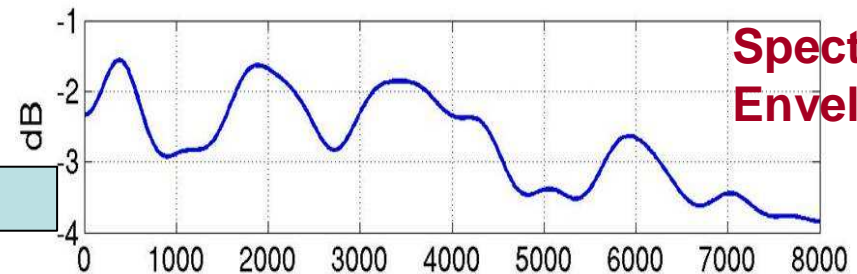


A pseudo-frequency  
axis

Spectrum

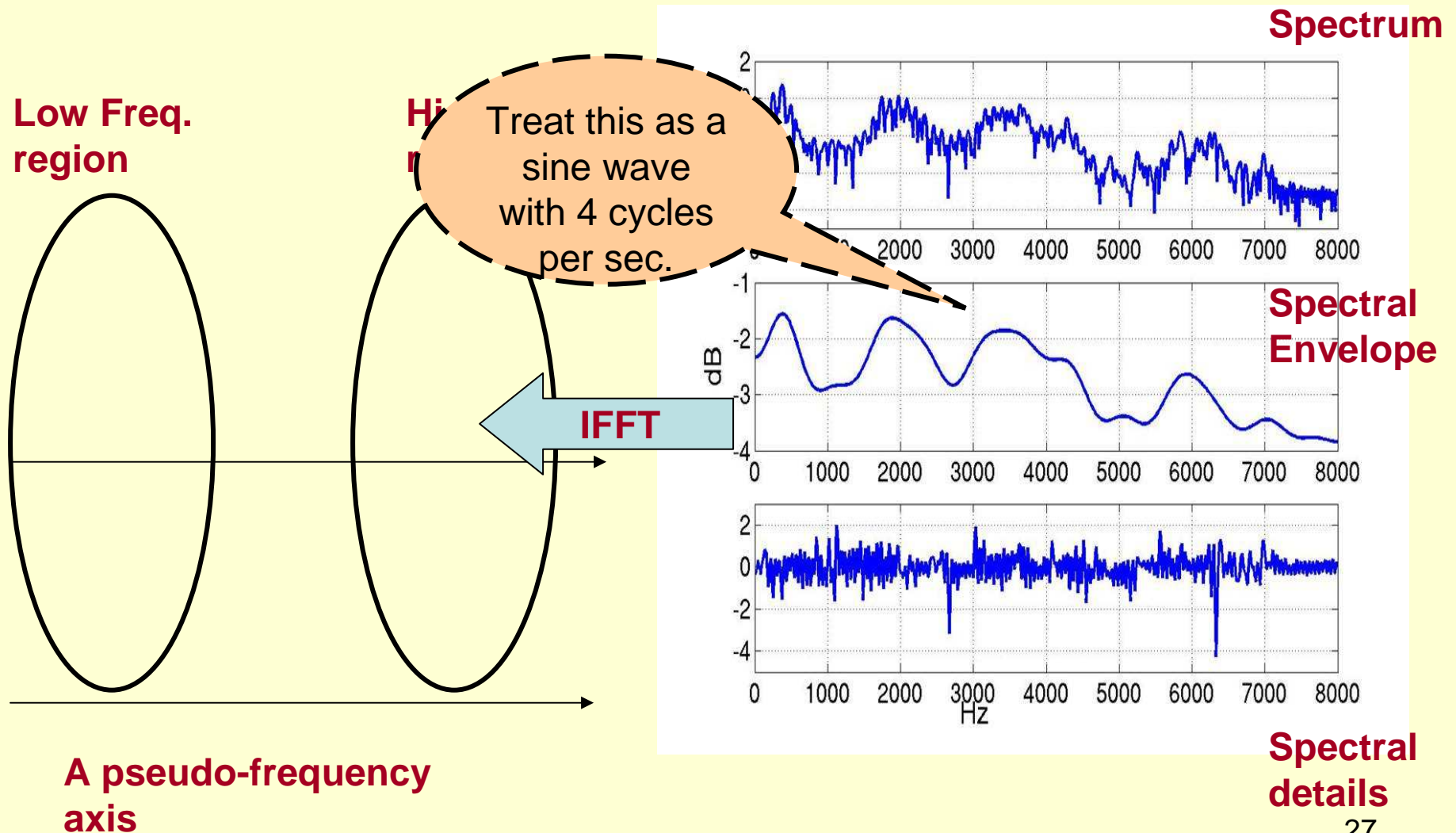


Spectral  
Envelope



Spectral  
details

# Play a Mathematical Trick



# a Mathematical Trick

Gives a peak  
at 4 Hz in  
frequency  
axis

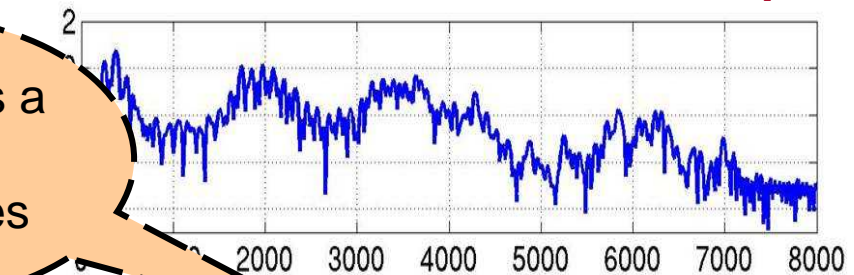
Low Freq.  
region

Treat this as a  
sine wave  
with 4 cycles  
per sec.

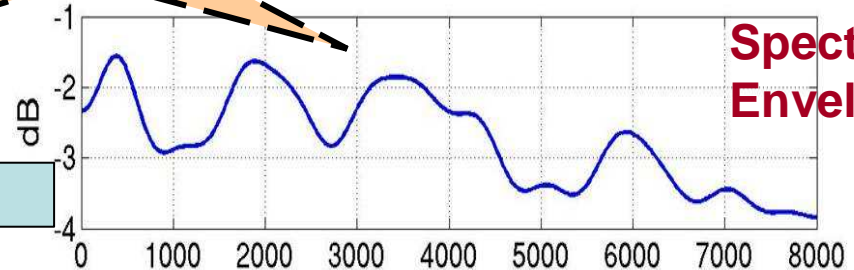
IFFT

A pseudo-frequency  
axis

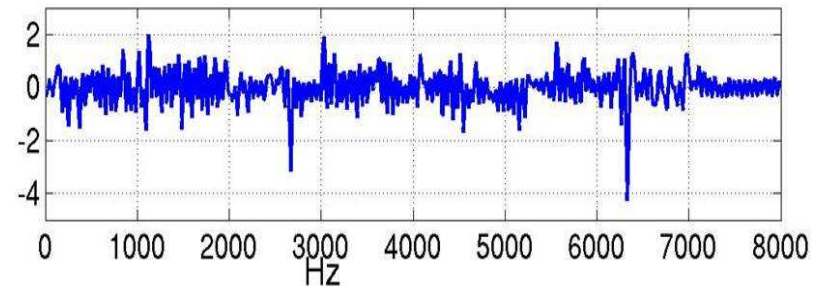
Spectrum



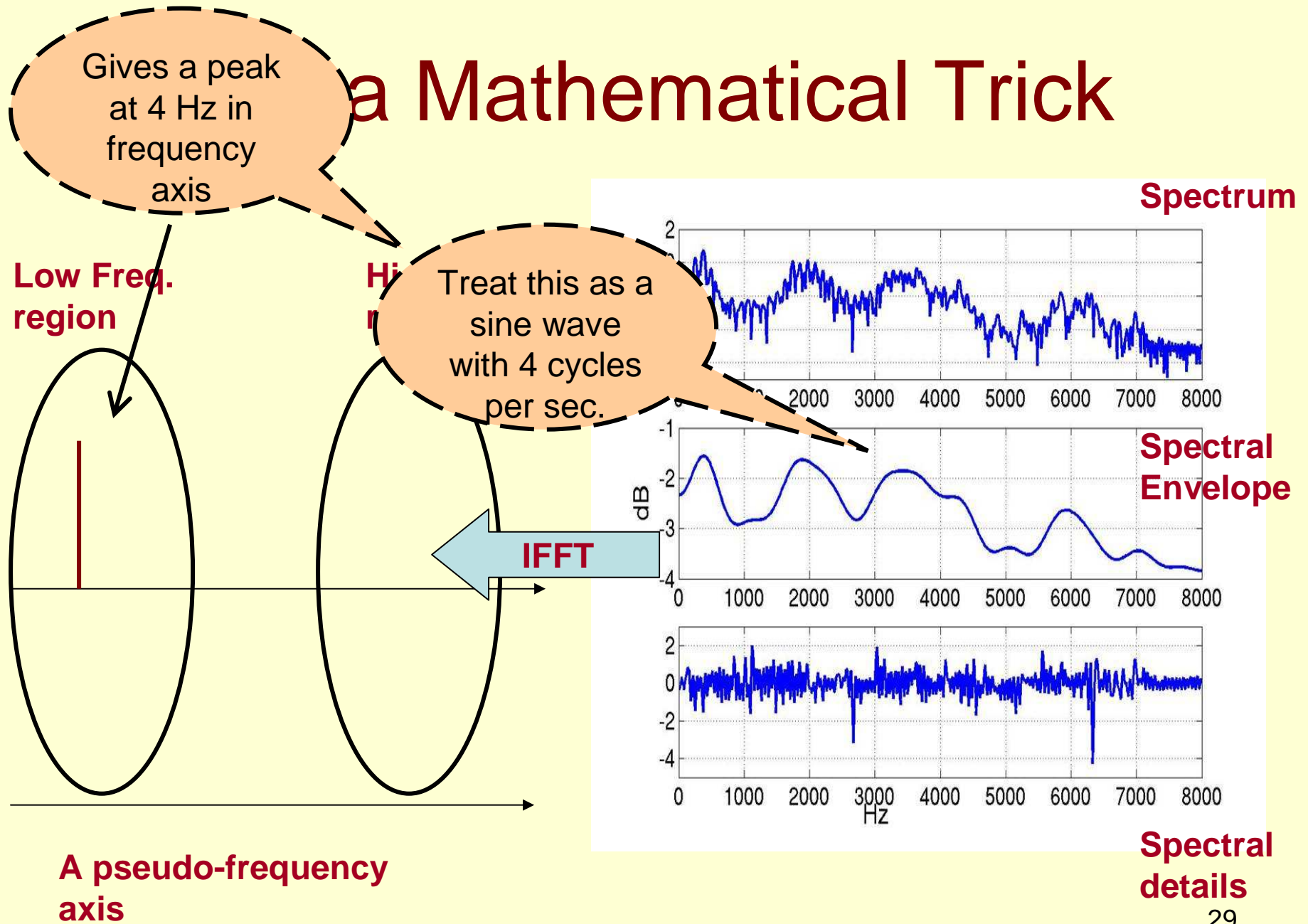
Spectral  
Envelope



Spectral  
details



# a Mathematical Trick

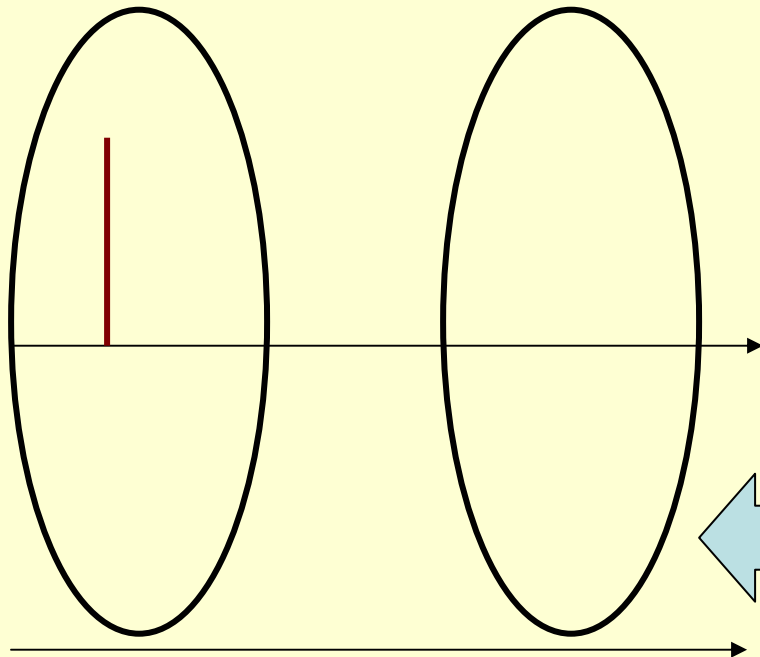




# Play a Mathematical Trick

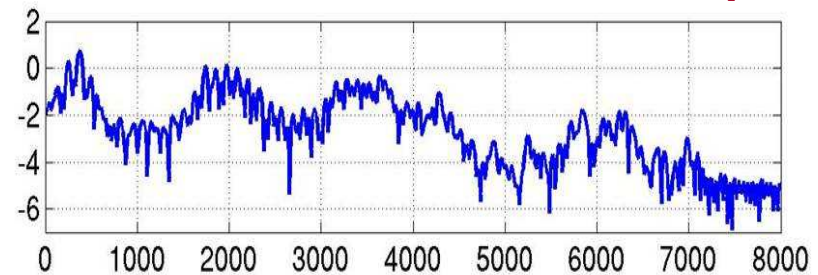
Low Freq.  
region

High Freq.  
region

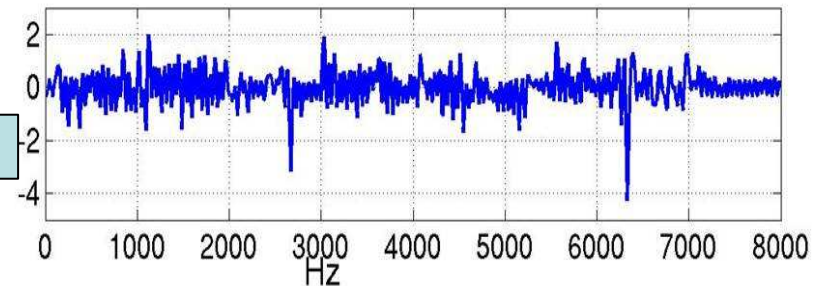
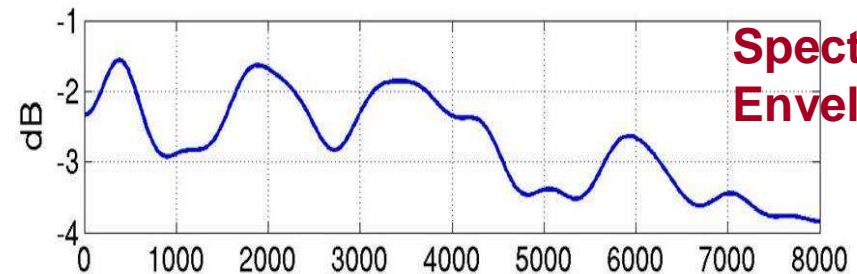


A pseudo-frequency  
axis

Spectrum



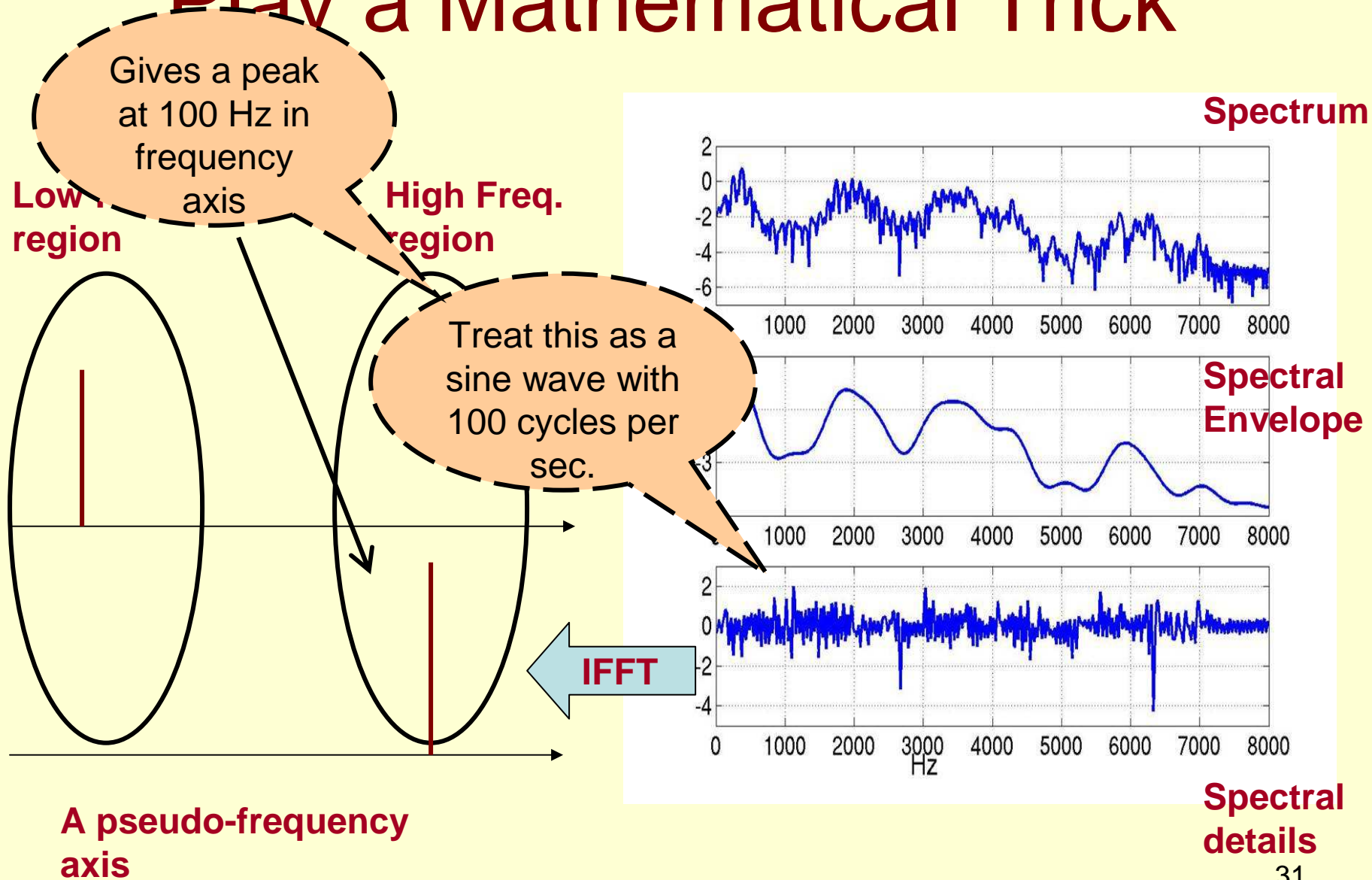
Spectral  
Envelope



Spectral  
details

30

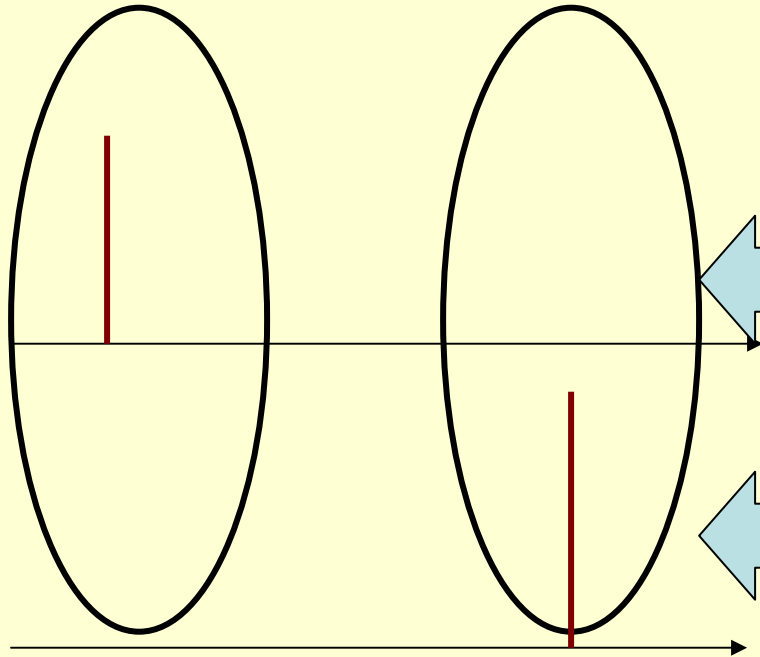
# Play a Mathematical Trick



# Play a Mathematical Trick

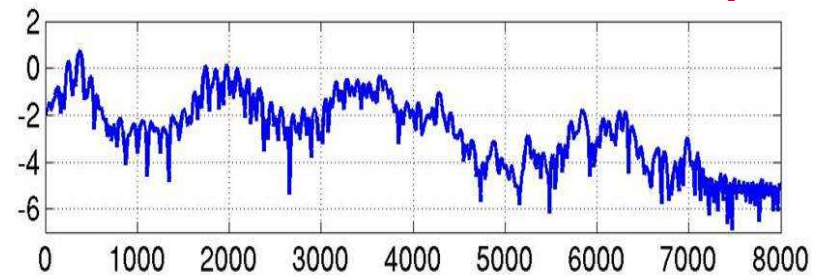
Low Freq.  
region

High Freq.  
region

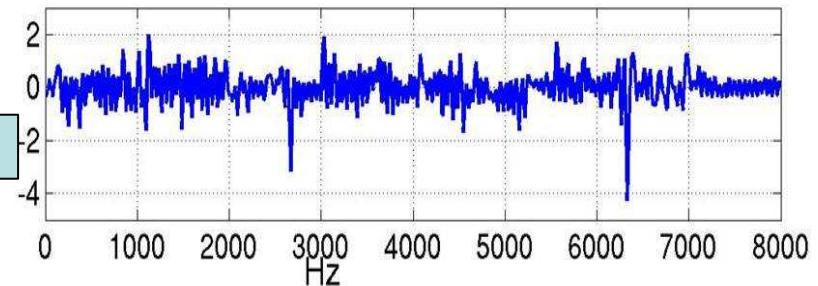
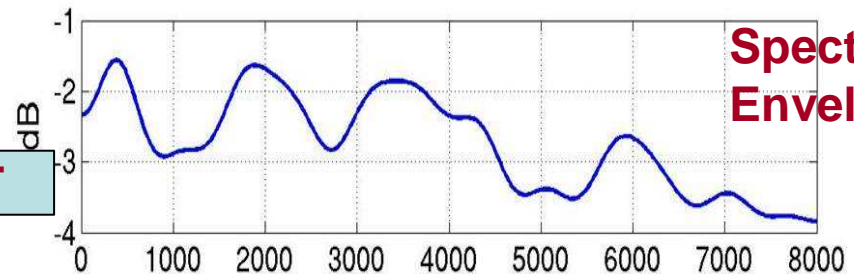


A pseudo-frequency  
axis

Spectrum



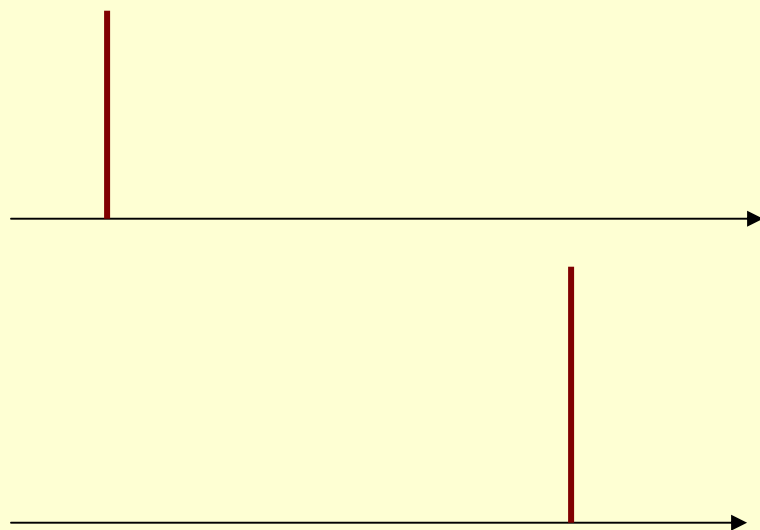
Spectral  
Envelope



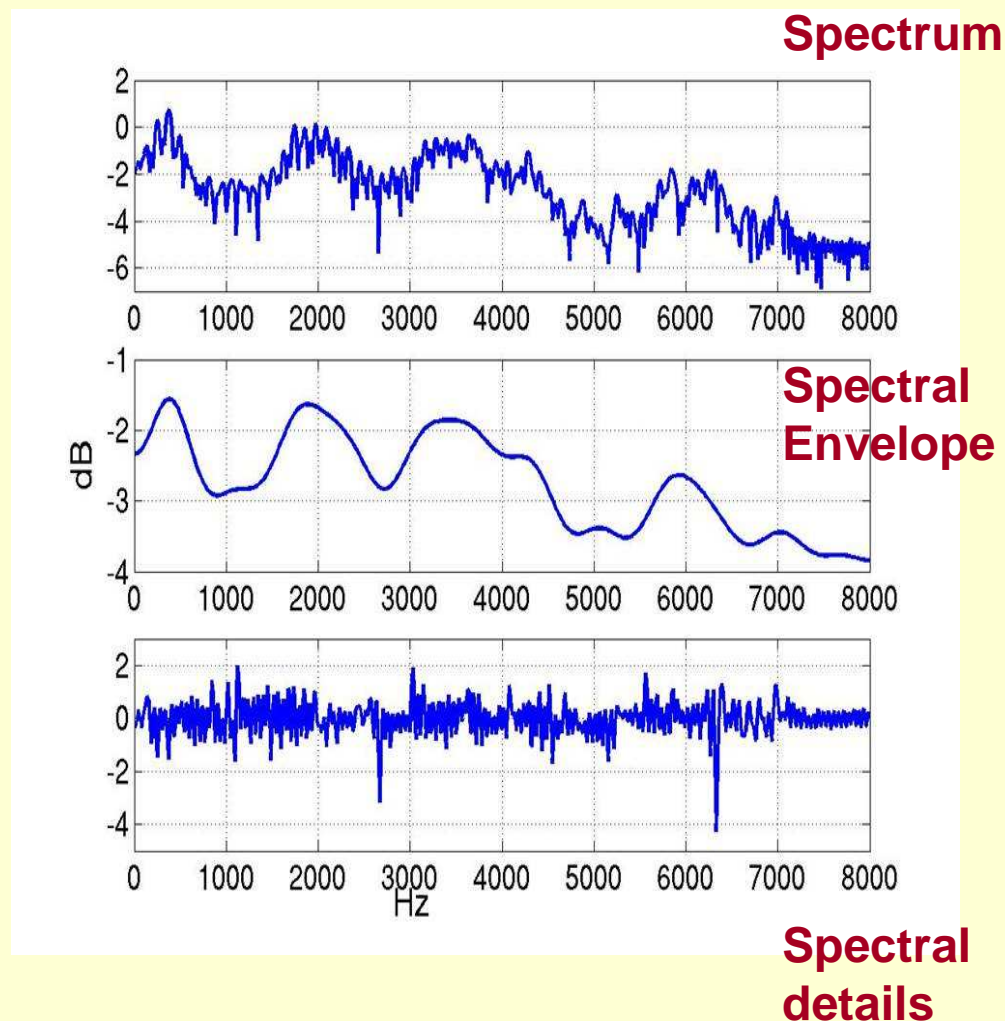
Spectral  
details



# Play a Mathematical Trick



**A pseudo-frequency axis**

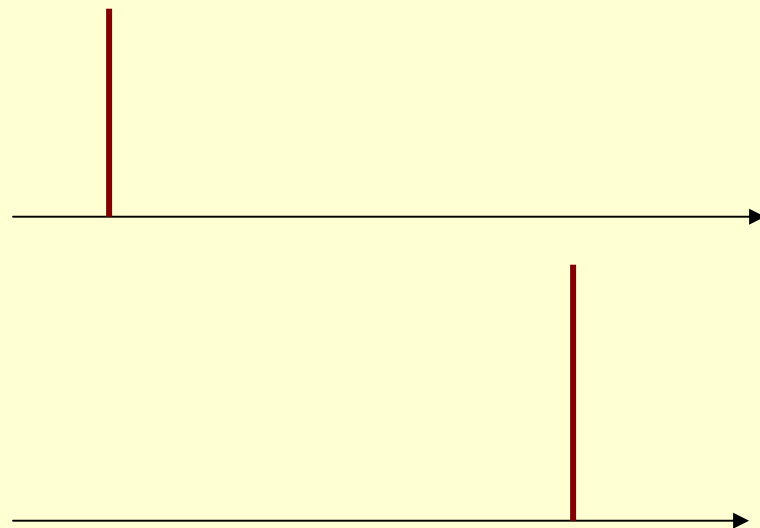


**Spectrum**

**Spectral Envelope**

**Spectral details**

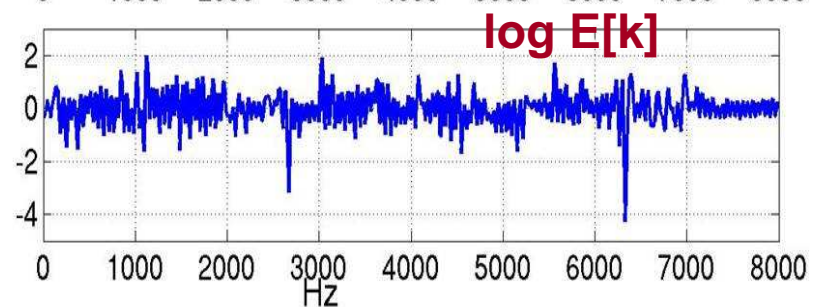
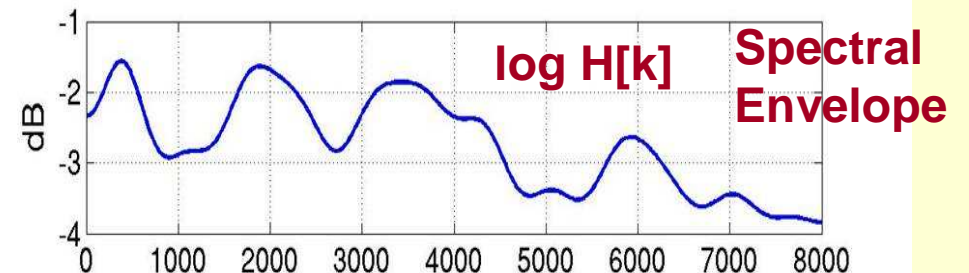
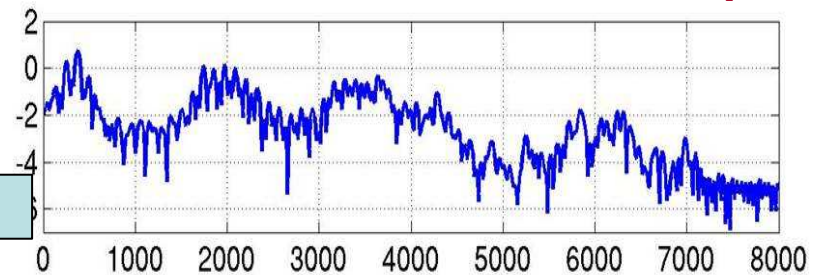
# Play a Mathematical Trick



A pseudo-frequency  
axis

IFFT

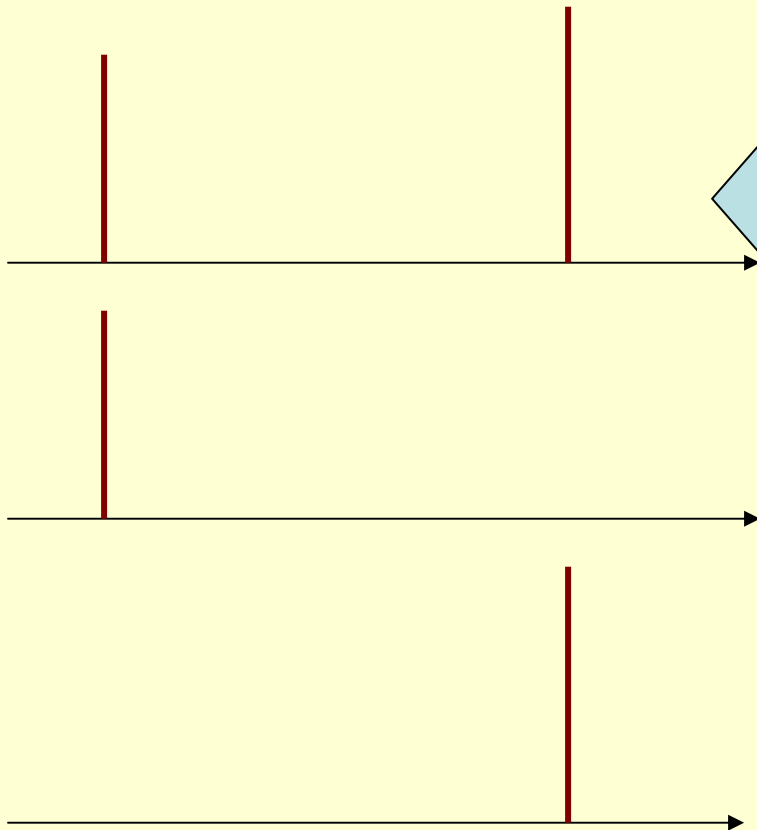
$$\log X[k] = \log H[k] + \log E[k] \quad \text{Spectrum}$$



Spectral  
details

# Play a Mathematical Trick

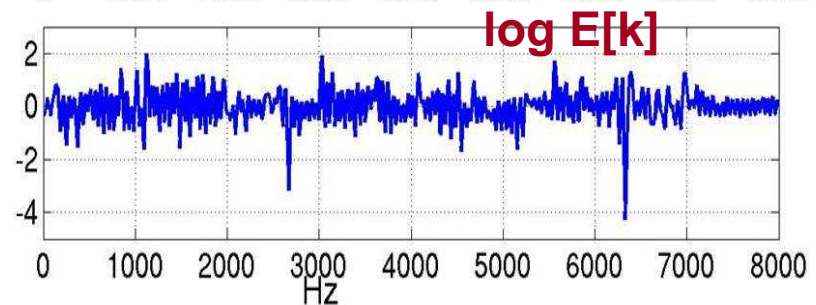
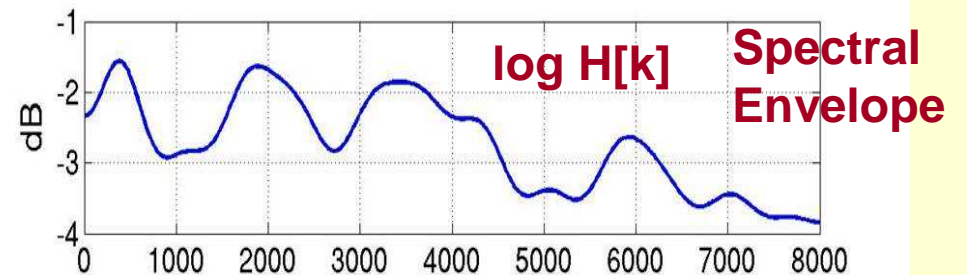
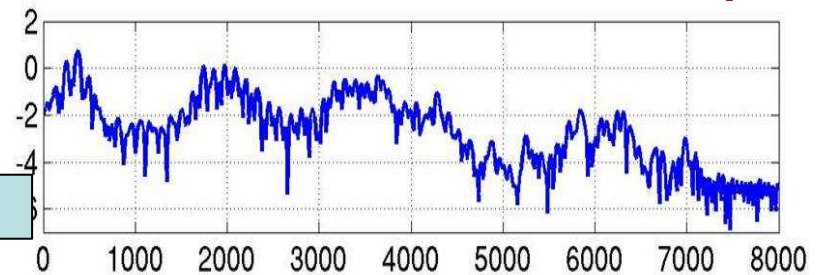
$$x[k] = h[k] + e[k]$$



A pseudo-frequency  
axis

IFFT

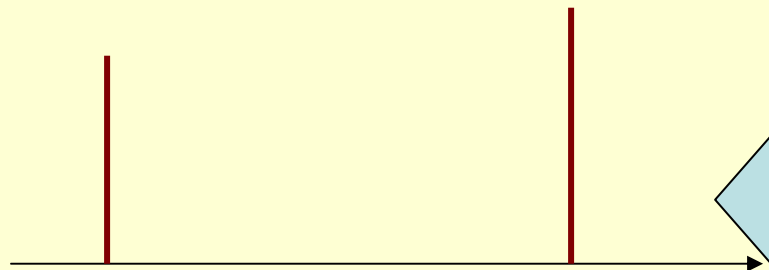
$$\log X[k] = \log H[k] + \log E[k] \quad \text{Spectrum}$$



Spectral  
details

# Play a Mathematical Trick

$$x[k] = h[k] + e[k]$$

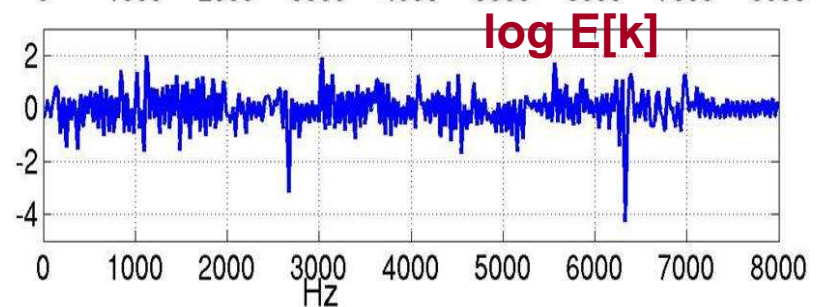
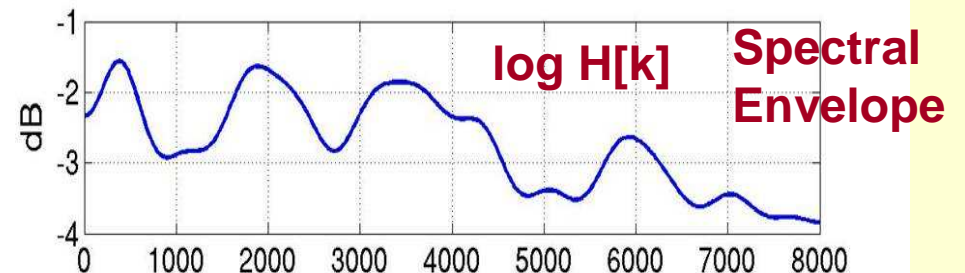
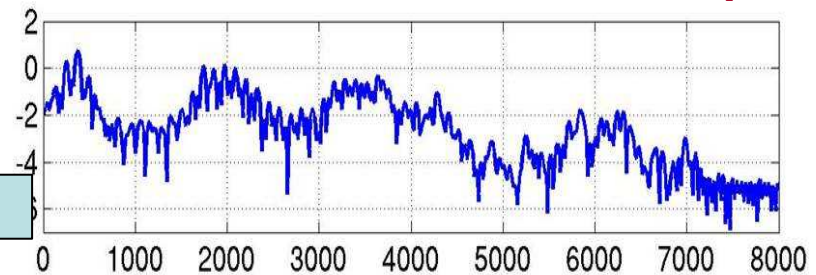


IFFT

In practice all you have access to only  $\log X[k]$  and hence you can obtain  $x[k]$

A pseudo-frequency axis

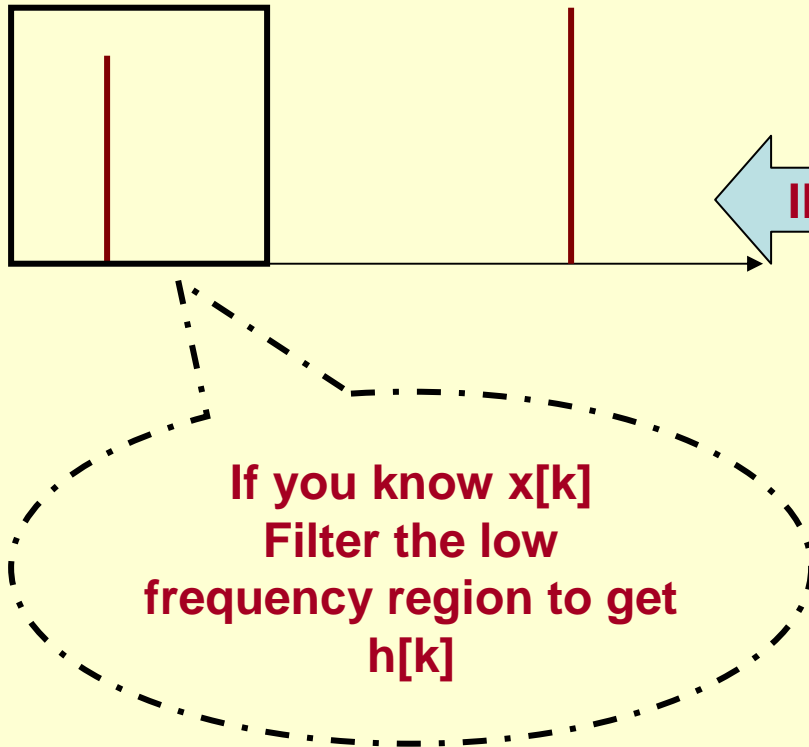
$$\log X[k] = \log H[k] + \log E[k] \quad \text{Spectrum}$$



Spectral details

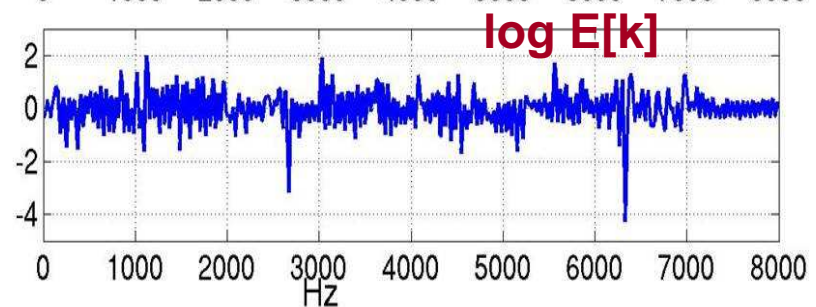
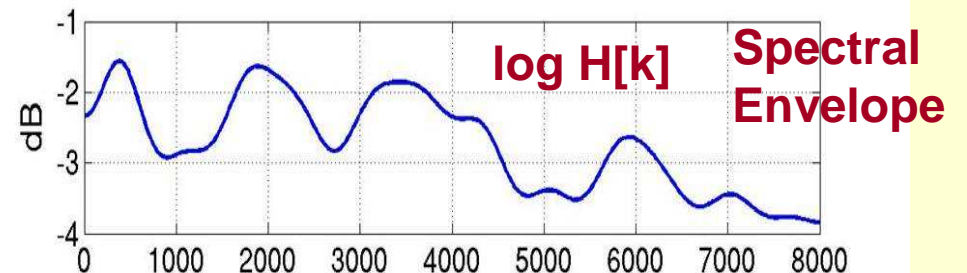
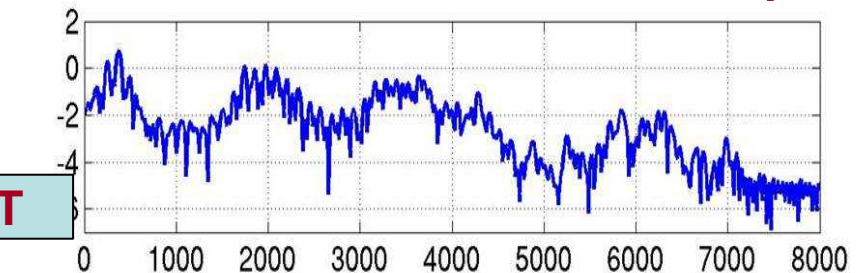
# Play a Mathematical Trick

$$x[k] = h[k] + e[k]$$



A pseudo-frequency  
axis

$$\log X[k] = \log H[k] + \log E[k] \quad \text{Spectrum}$$

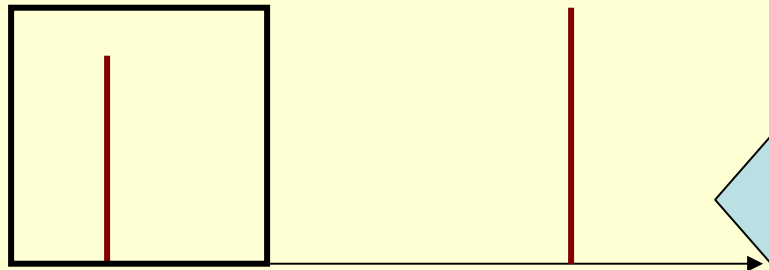


Spectral  
details



# Play a Mathematical Trick

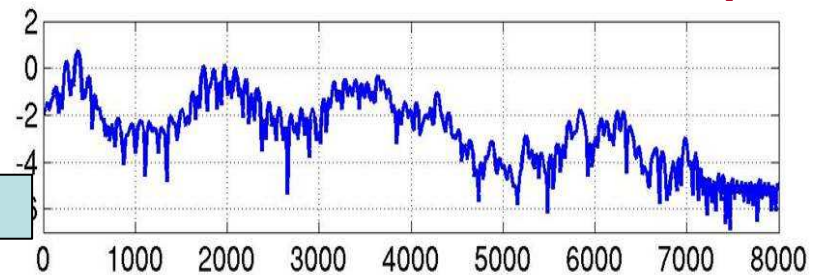
$$x[k] = h[k] + e[k]$$



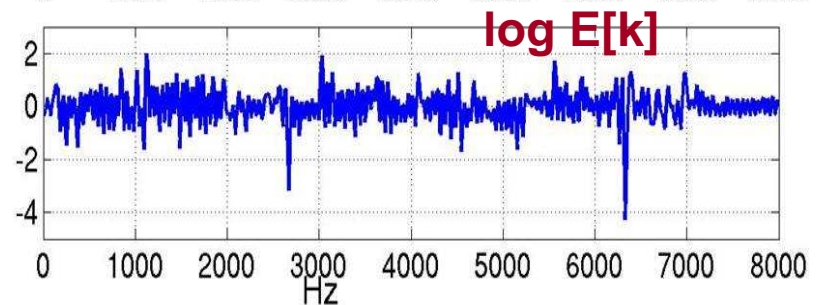
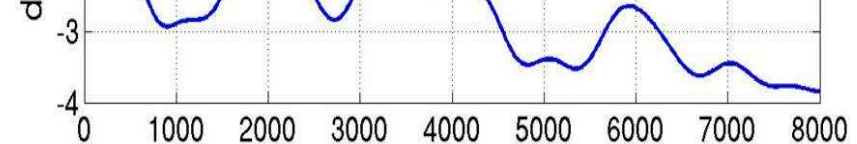
A pseudo-frequency axis

- $x[k]$  is referred to as **Cepstrum**
- $h[k]$  is obtained by considering the low frequency region of  $x[k]$ .
- $h[k]$  represents the spectral envelope and is widely used as feature for speech recognition

$$\log X[k] = \log H[k] + \log E[k] \quad \text{Spectrum}$$



$$\log H[k] \quad \text{Spectral Envelope}$$



Spectral details

# Cepstral Analysis

$$X[k] = H[k] E[k]$$

$$\| X[k] \| = \| H[k] \| \| E[k] \|^2$$

$\| \cdot \|$  –denotes magnitude

Take Log on both sides

$$\log \| X[k] \| = \log \| H[k] \| + \log \| E[k] \|^2$$

Taking inverse FFT on both sides

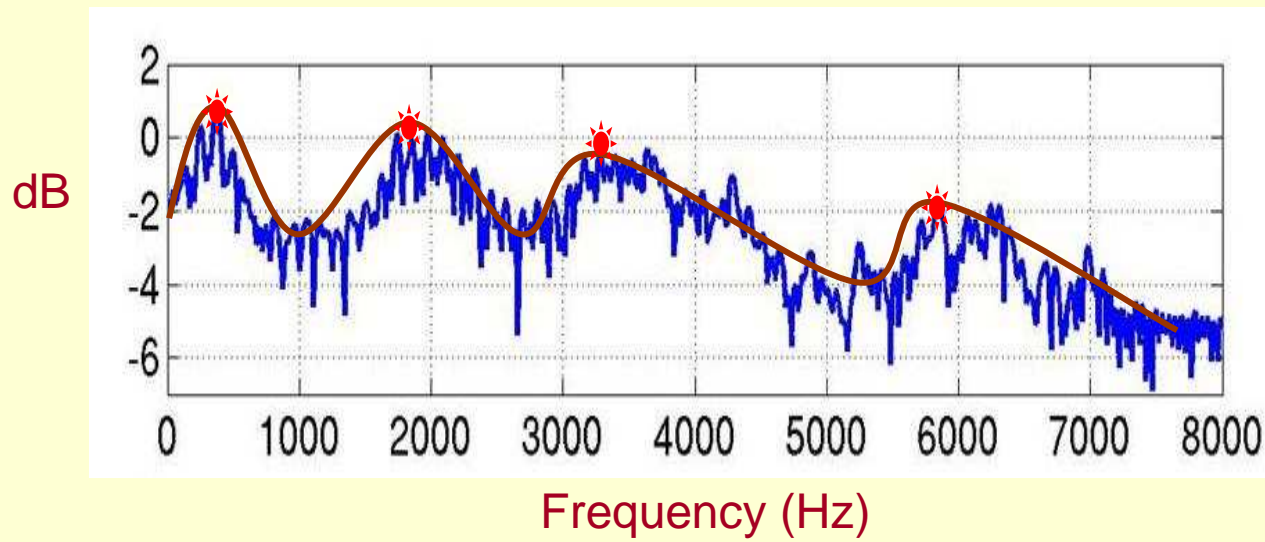
$$x[k] = h[k] + e[k]$$

# Mel-Frequency Analysis



# Review: What we did

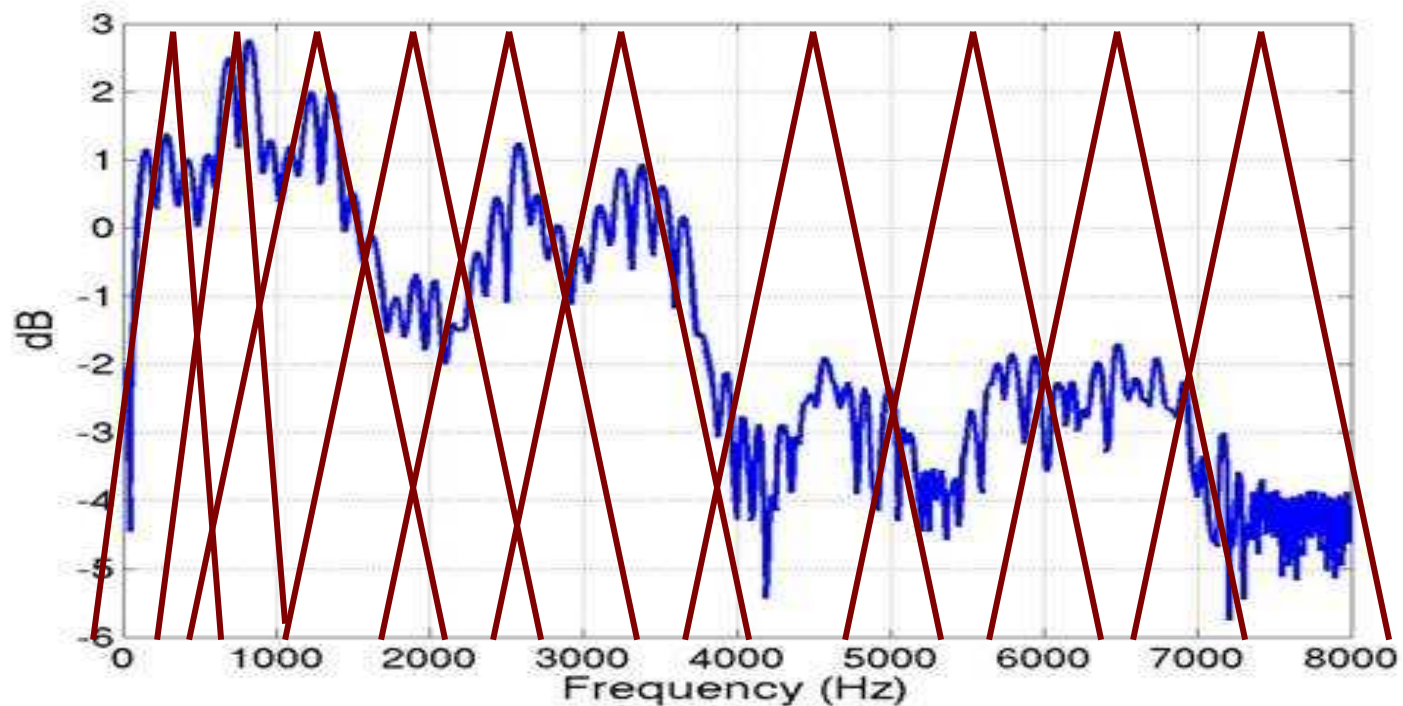
- We captured spectral envelope (curve connecting all formants)
- BUT: Perceptual experiments say human ear concentrates on certain regions rather than using whole of the spectral envelope....



# Mel-Frequency Analysis

- Mel-Frequency analysis of speech is based on human perception experiments
- It is observed that human ear acts as filter
  - It concentrates on only certain frequency components
- These filters are non-uniformly spaced on the frequency axis
  - More filters in the low frequency regions
  - Less no. of filters in high frequency regions

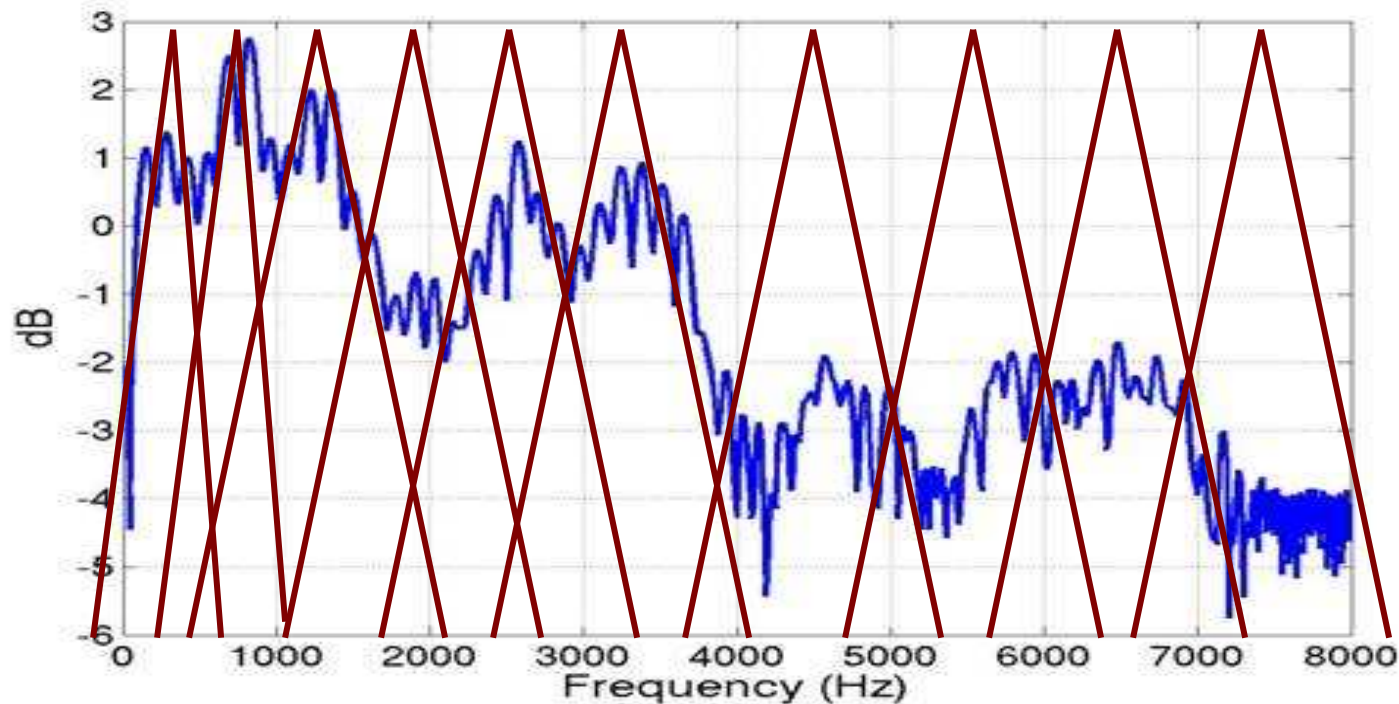
# Mel-Frequency Filters



# Mel-Frequency Filters

More no. of filters in low  
freq. region

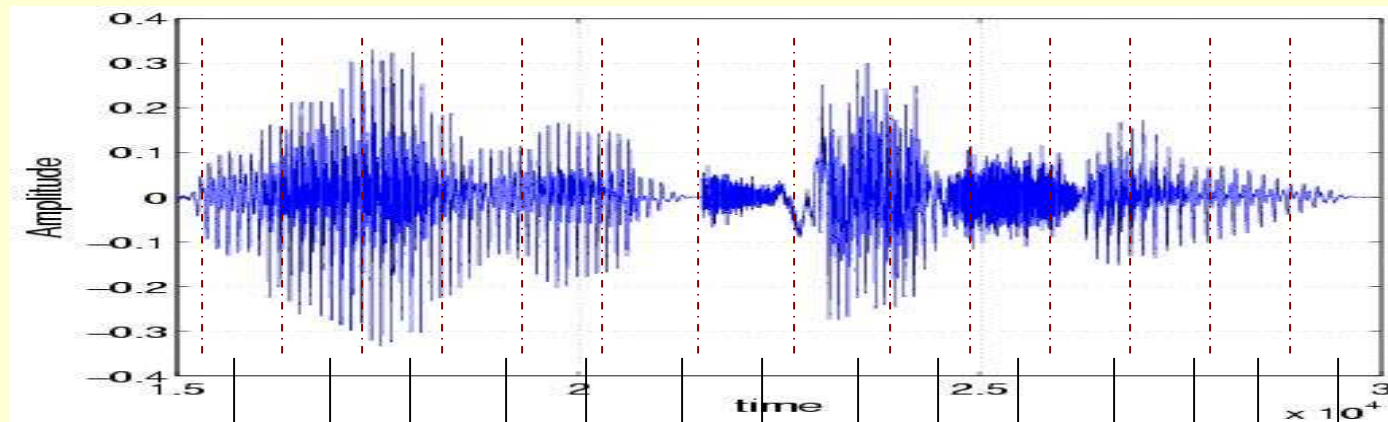
Lesser no. of filters in  
high freq. region



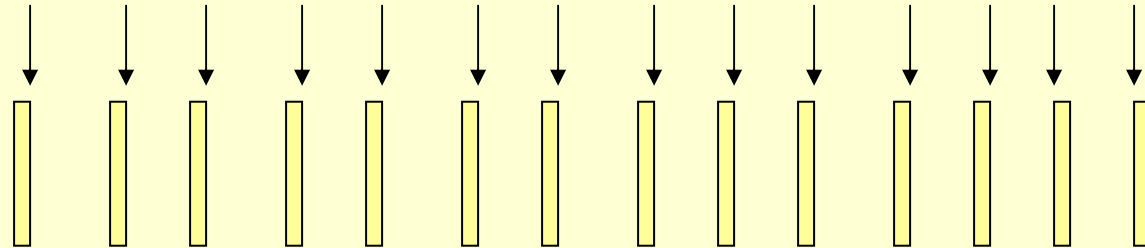
# Mel-Frequency Cepstral Coefficients (MFCC)

- Spectrum  $\rightarrow$  Mel-Filters  $\rightarrow$  Mel-Spectrum
- Say  $\log X[k] = \log (\text{Mel-Spectrum})$
- NOW perform Cepstral analysis on  $\log X[k]$ 
  - $\log X[k] = \log H[k] + \log E[k]$
  - Taking IFFT
  - $x[k] = h[k] + e[k]$
- Cepstral coefficients  $h[k]$  obtained for Mel-spectrum are referred to as Mel-Frequency Cepstral Coefficients often denoted by \*MFCC\*

# Speech signal represented as a sequence of spectral vectors



FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT



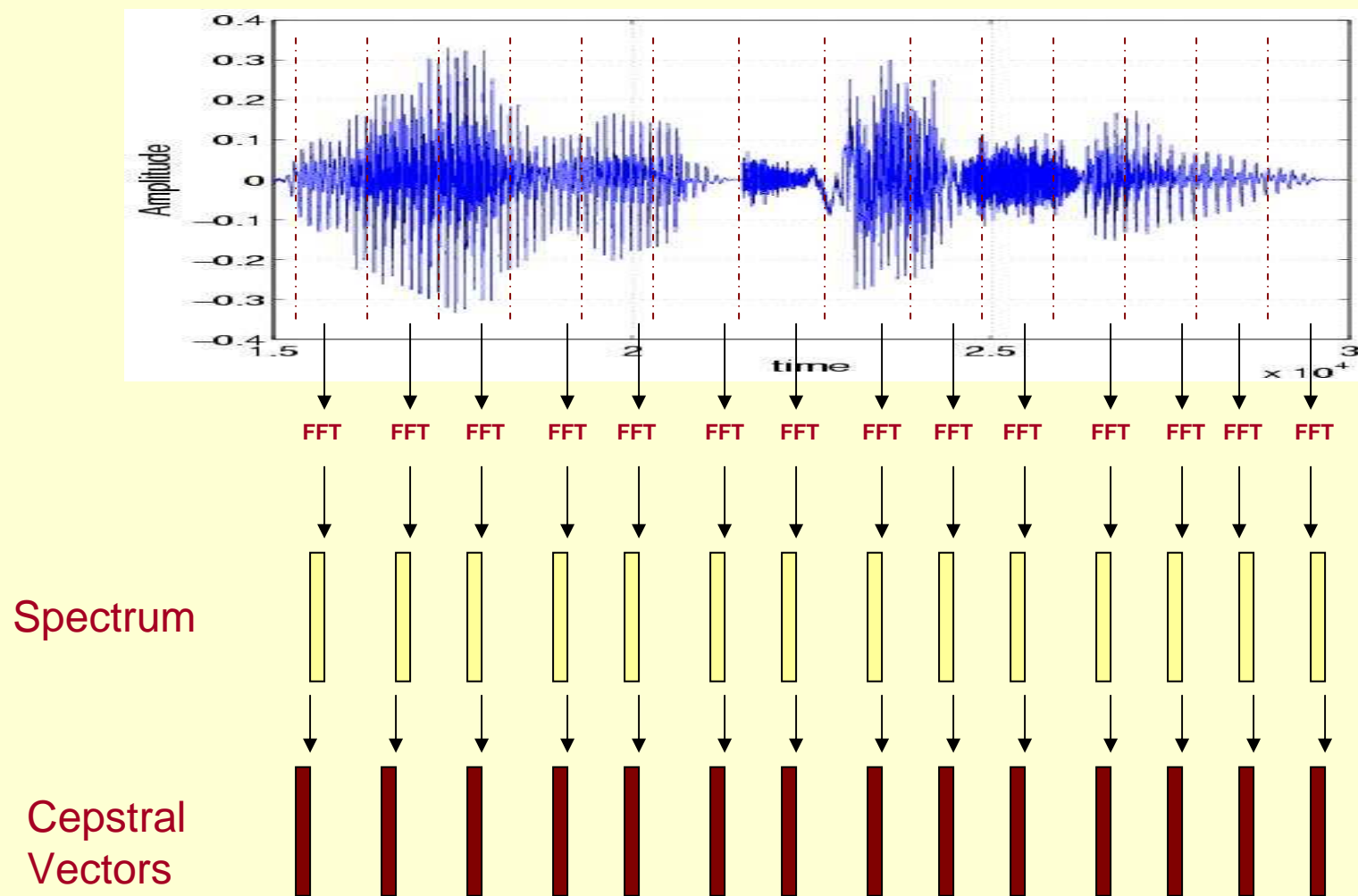
Spectrum

Mel-Filters

Cepstral Analy.



## Speech signal represented as a sequence of CEPSTRAL vectors





# Why we are going to use MFCC

- Speech synthesis
  - Used for joining two speech segments S1 and S2
  - Represent S1 as a sequence of MFCC
  - Represent S2 as a sequence of MFCC
  - Join at the point where MFCCs of S1 and S2 have minimal Euclidean distance
- Used in speech recognition
  - MFCC are mostly used features in state-of-art speech recognition system

# Summary: Process of Feature Extraction

- Speech is analyzed over short analysis window
- For each short analysis window a spectrum is obtained using FFT
- Spectrum is passed through Mel-Filters to obtain Mel-Spectrum
- Cepstral analysis is performed on Mel-Spectrum to obtain Mel-Frequency Cepstral Coefficients
- Thus speech is represented as a sequence of Cepstral vectors
- It is these Cepstral vectors which are given to pattern classifiers for speech recognition purpose

# Additional Reading

- Chapter 6
  - Pg: 273 – 281
  - Pg: 304 – 311
  - Pg: 314 - 316