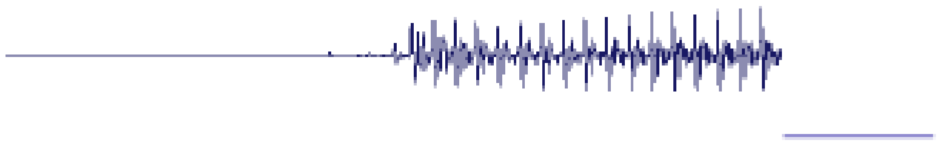
# Pretext tasks
# 4. CPC

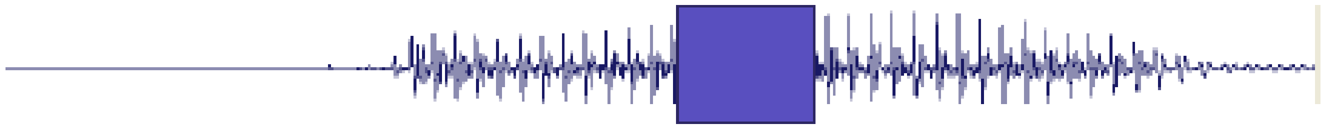
| | | | | |
|---|---|---|---|---|
| 1 2 | SELF-PREDICTION | INNATE RELATIONSHIP (Context-based) | 1. ROTATION<br>2. RELATIVE POSITION | IMAGE |
| 3 | CONTRASTIVE LEARNING | INTER-SAMPLE CLASSIFICATION | 1. Instance Discrimination<br>2. SimCLR [Contrastive Loss]<br>3. Theory – Guarantees / Bounds | IMAGE |
| 4 | CONTRASTIVE LEARNING | INTER-SAMPLE CLASSIFICATION | Contrastive Predictive Coding (CPC), [NCE, InfoNCE Loss] | AUDIO/ SPEECH |
| 5 | SELF-PREDICTION | GENERATIVE (VAE) | 1. AE – Variational Bayes<br>2. VQ-VAE + AR | IMAGE<br>AUDIO/ SPEECH |
| 6 | SELF-PREDICTION | GENERATIVE (AR) | 1. AR-LM – GPT<br>2. Masked-LM – BERT | LANGUAGE |
| 7 | SELF-PREDICTION | MASKED-GEN (Masked LM for ASR) | 1. Wav2Vec / 2.0<br>2. HuBERT | AUDIO/ SPEECH |

# Learning with or without supervision – speech and audio

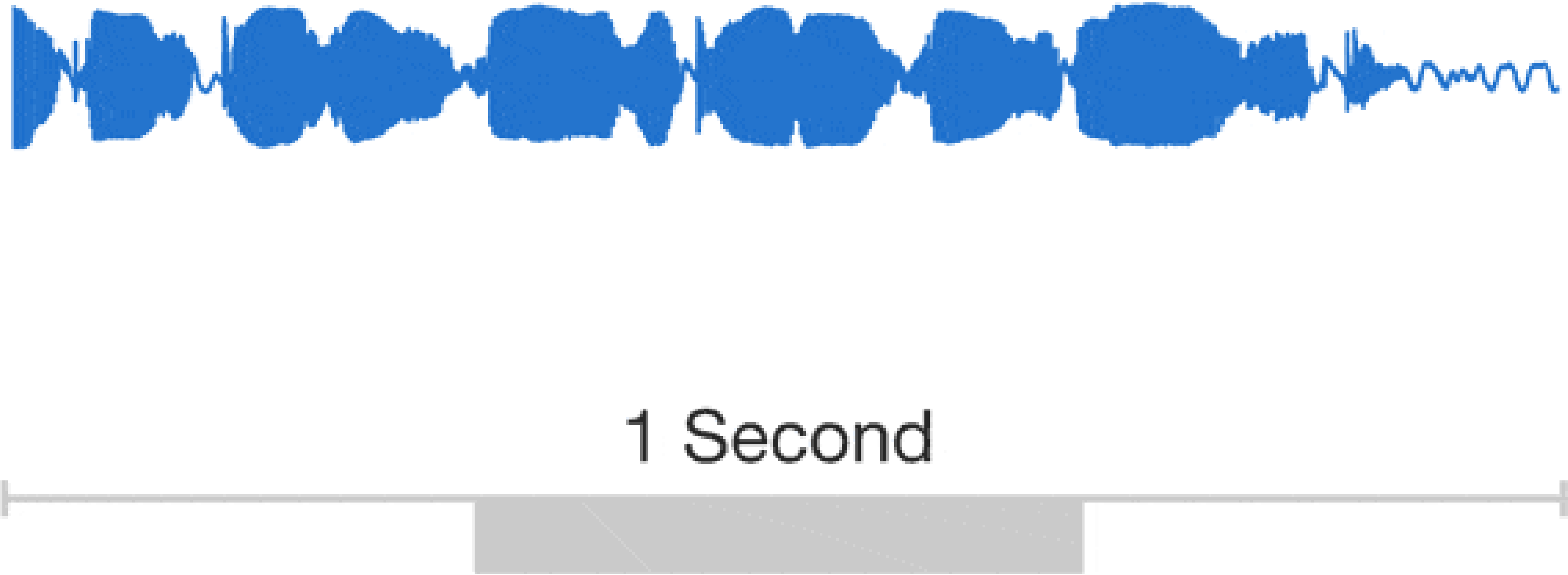
- Next frame prediction



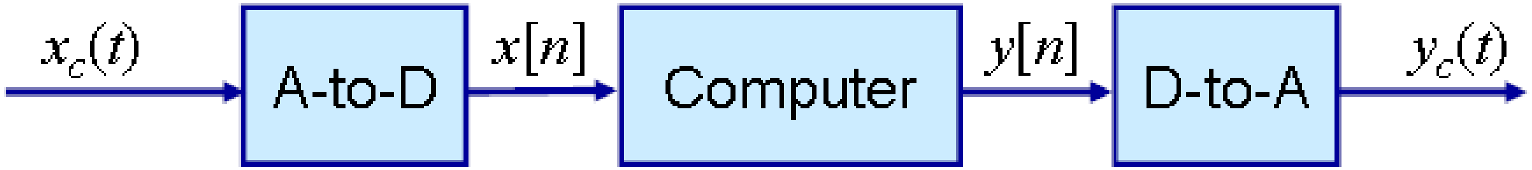
- Masked prediction



# Speech Waveform Basics


1 Second

# Digital Processing of Analog Signals




- **A-to-D conversion:** bandwidth control, sampling and quantization
- **Computational processing:** implemented on computers or ASICs with finite-precision arithmetic
  - **basic numerical processing:** add, subtract, multiply (scaling, amplification, attenuation), mute, …
  - **algorithmic numerical processing:** convolution or linear filtering, non-linear filtering (e.g., median filtering), difference equations, DFT, inverse filtering, MAX/MIN, …
- **D-to-A conversion:** re-quantification* and filtering (or interpolation) for reconstruction

# Discrete-Time Signals

- A sequence of numbers
- Mathematical representation:

$$x = \{x[n]\}, \qquad -\infty < n < \infty$$

- Sampled from an analog signal, $x_a(t)$, at time $t = nT$,

$$x[n] = x_a(nT), \qquad -\infty < n < \infty$$

- $T$ is called the **sampling period,** and its reciprocal, $F_S = 1/T$, is called the **sampling frequency**

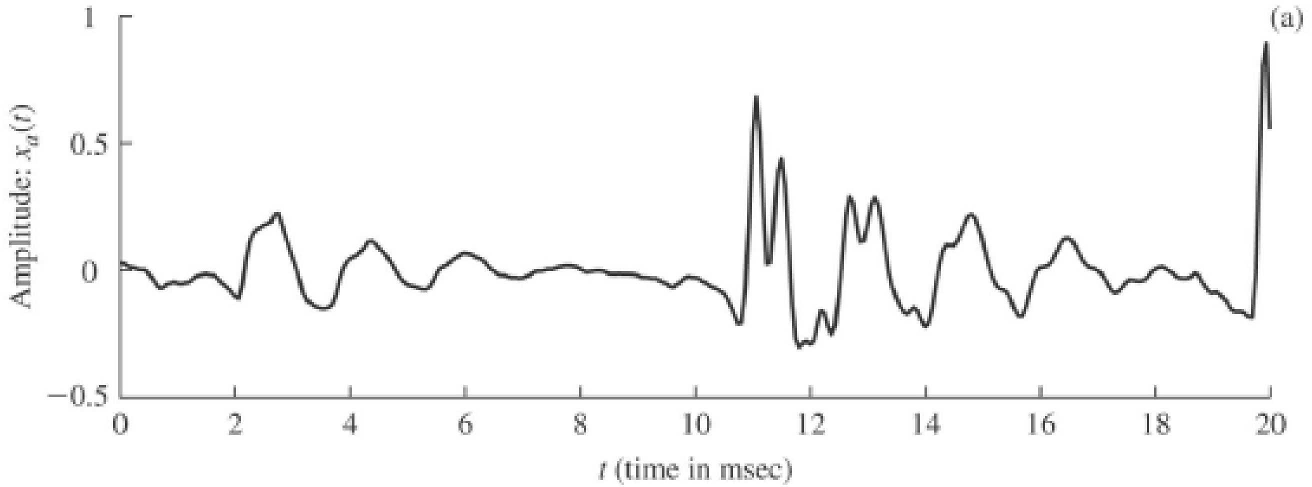
$$F_S = 8000 \text{ Hz} \quad \leftrightarrow \quad T = 1/8000 = 125\ \mu\text{sec}$$

$$F_S = 10000 \text{ Hz} \quad \leftrightarrow \quad T = 1/10000 = 100\ \mu\text{sec}$$

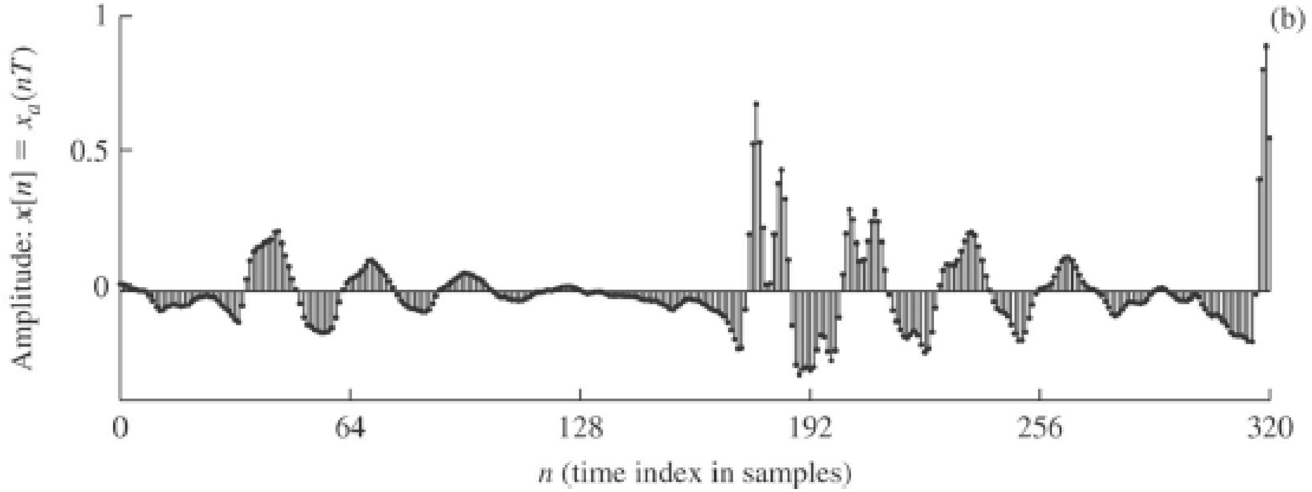
$$F_S = 16000 \text{ Hz} \quad \leftrightarrow \quad T = 1/16000 = 62.5\ \mu\text{sec}$$

$$F_S = 20000 \text{ Hz} \quad \leftrightarrow \quad T = 1/20000 = 50\ \mu\text{sec}$$

# Speech Waveform Display

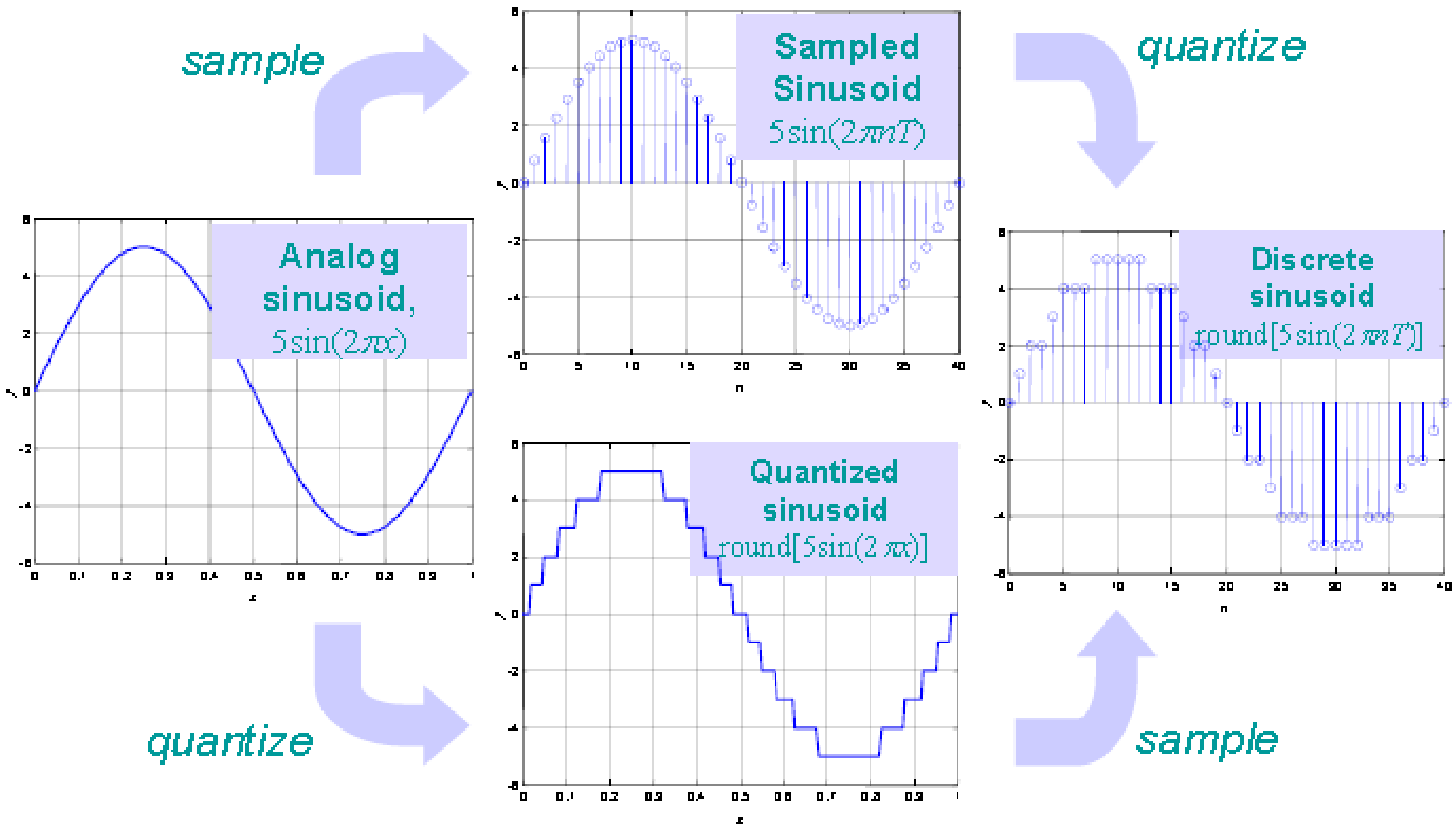



plot( );

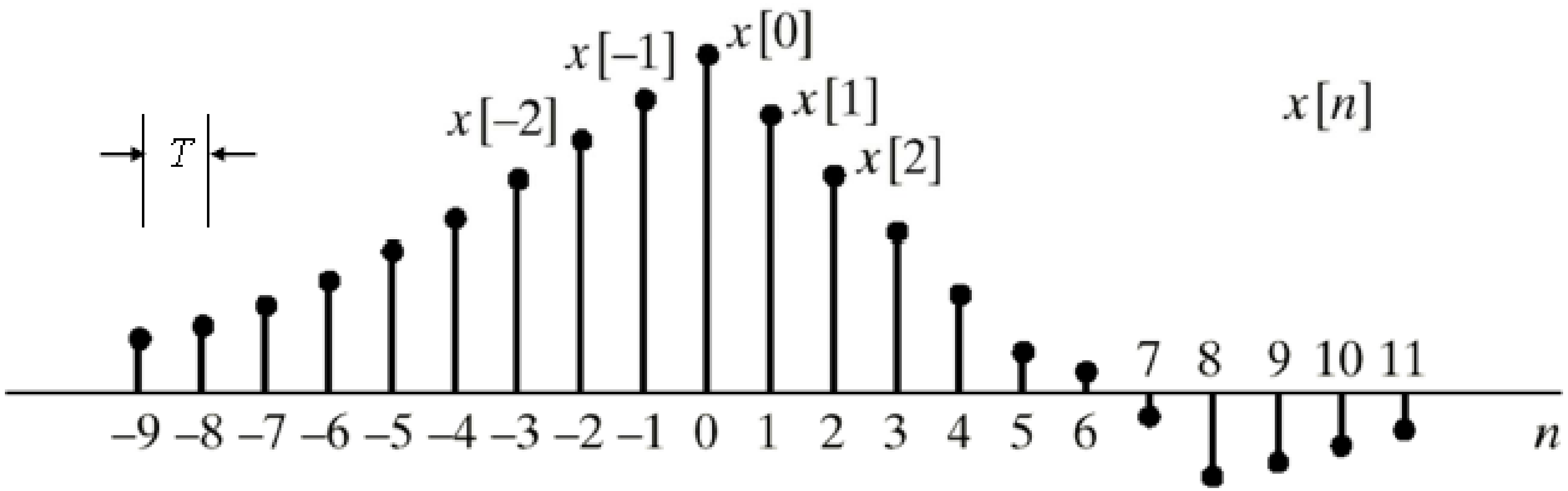



stem( );

# Discrete Signals

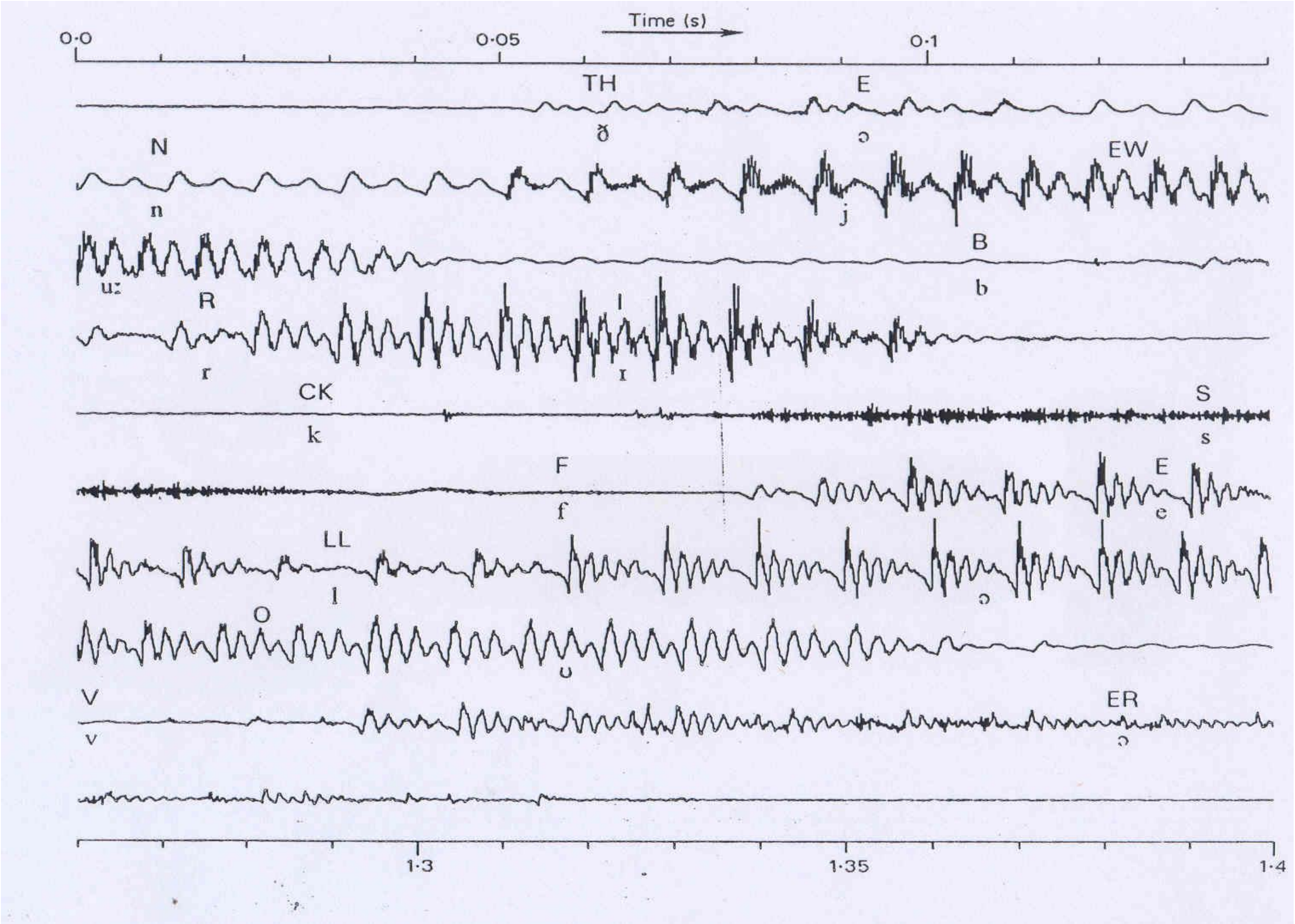


# Discrete-Time (DT) Signals are Sequences




- $x[n]$ denotes the "sequence value at 'time' $n$"
- Sources of sequences:
  - Sampling a continuous-time signal
    $x[n] = x_c(nT) = x_c(t)|_{t=nT}$
  - Mathematical formulas – generative system
    e.g., $x[n] = 0.3 \bullet x[n\text{-}1] \text{-}1; \quad x[0] = 40$

## Speech waveform example




THE NEW BRICKS FELL OVER

# Speech spectrogram example

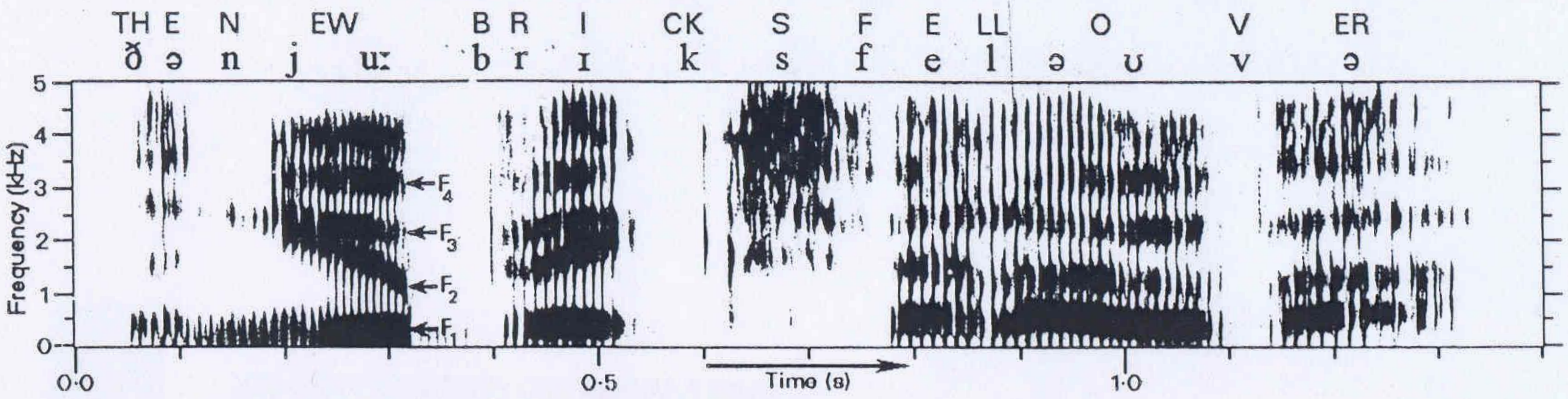
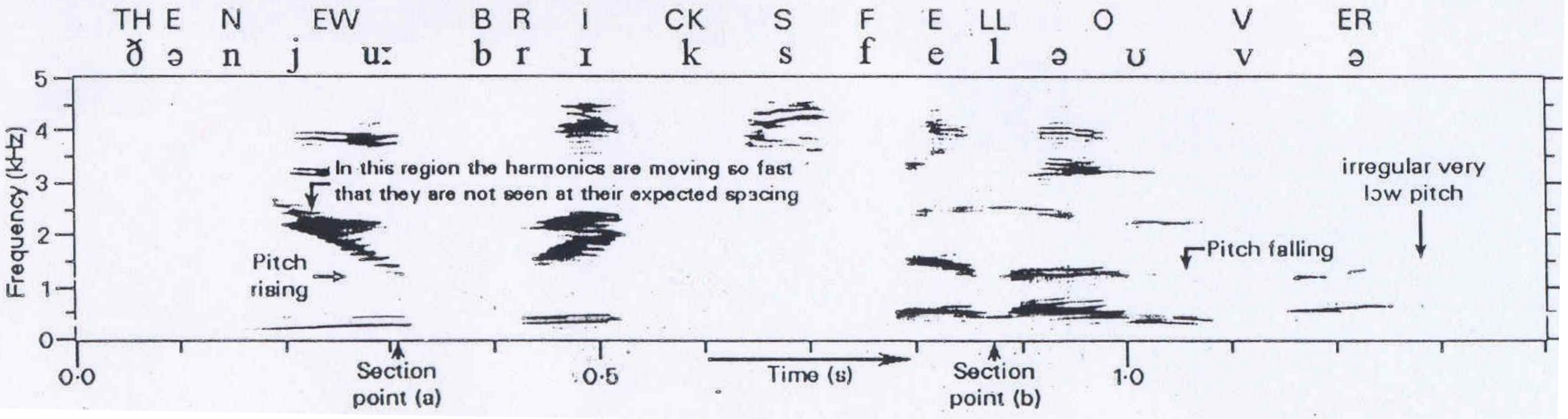
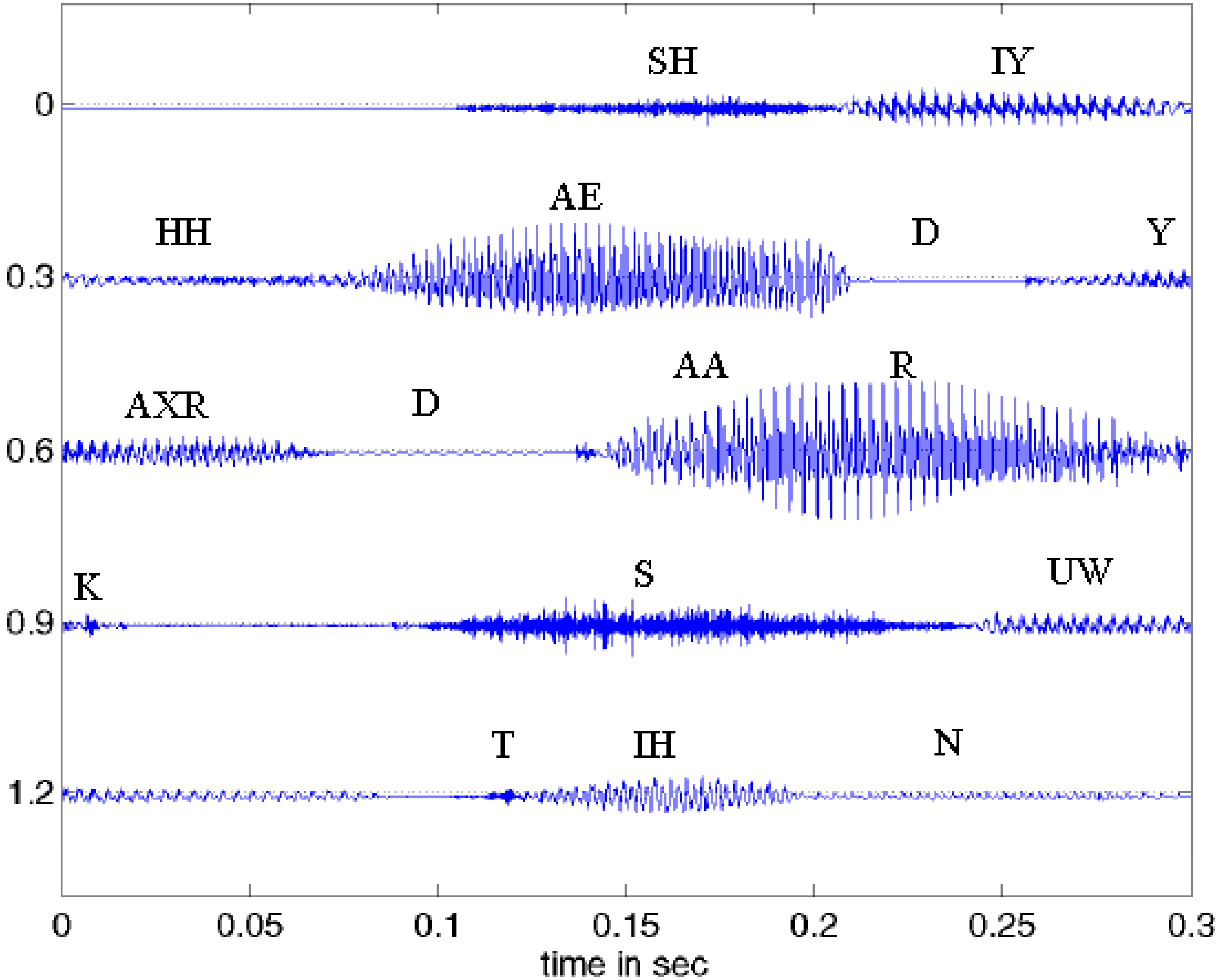


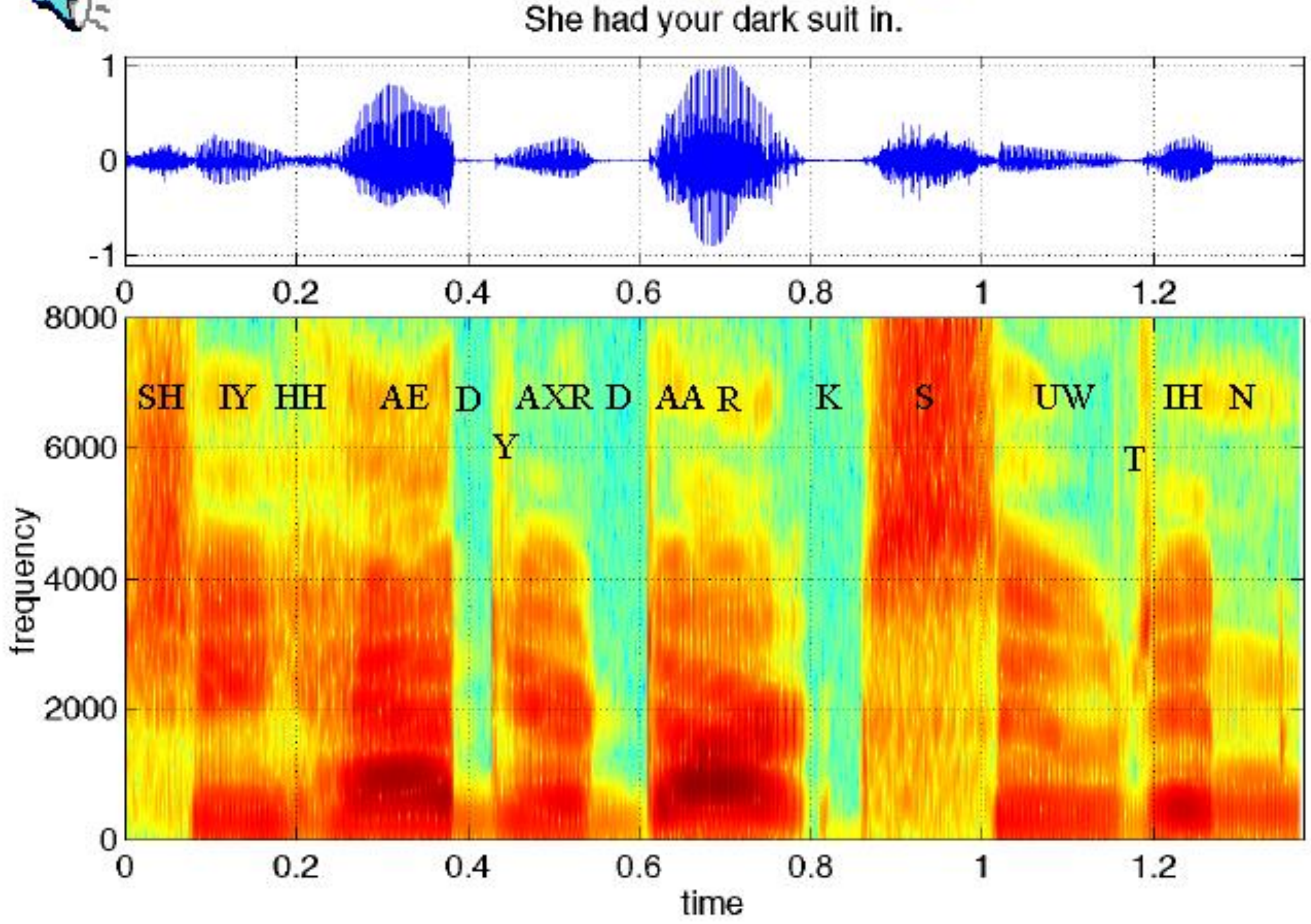

**Figure 2.10** Wide-band spectrogram of the speech waveform shown in Figure 2.9. The dynamic range of the grey scale in the display is 50 dB, so very weak sounds are clearly visible.



# She had your dark suit in…

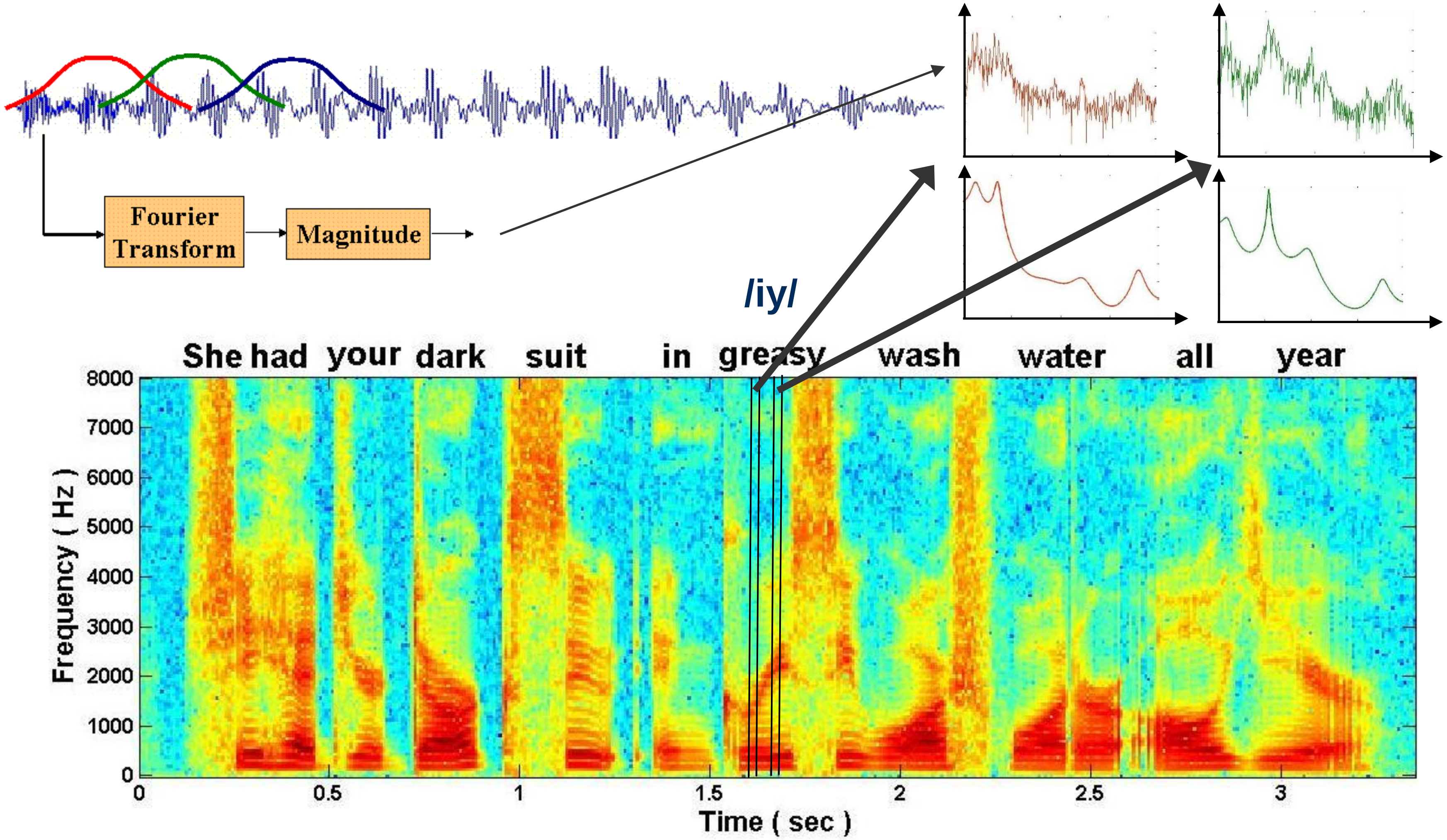


# “Wideband” Spectrogram



## Spectrogram

- Speech is a continuous evolution of the vocal tract
- Spectrogram shows time-frequency evolution
- Represented as a time-series of short-time spectra



# Spectrum Basics

FROM LAY's "LINEAR ALGEBRA" BOOK



# Fourier Series (Calculus required)

Continuous functions are often approximated by linear combinations of sine and cosine functions. For instance, a continuous function might represent a sound wave, an electric signal of some type, or the movement of a vibrating mechanical system.

For simplicity, we consider functions on $0 \le t \le 2\pi$. It turns out that any function in $C[0, 2\pi]$ can be approximated as closely as desired by a function of the form

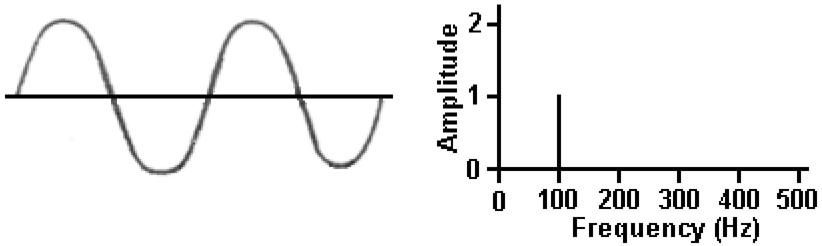
$$\frac{a_0}{2} + a_1 \cos t + \cdots + a_n \cos nt + b_1 \sin t + \cdots + b_n \sin nt \tag{4}$$

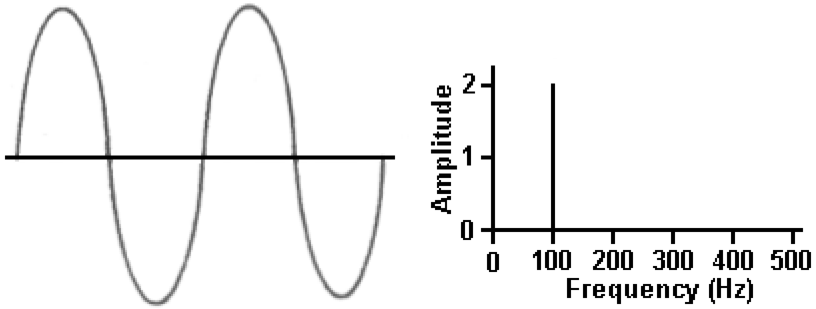
for a sufficiently large value of $n$. The function (4) is called a **trigonometric polynomial**. If $a_n$ and $b_n$ are not both zero, the polynomial is said to be of **order** $n$. The connection between trigonometric polynomials and other functions in $C[0, 2\pi]$ depends on the fact that for any $n \ge 1$, the set

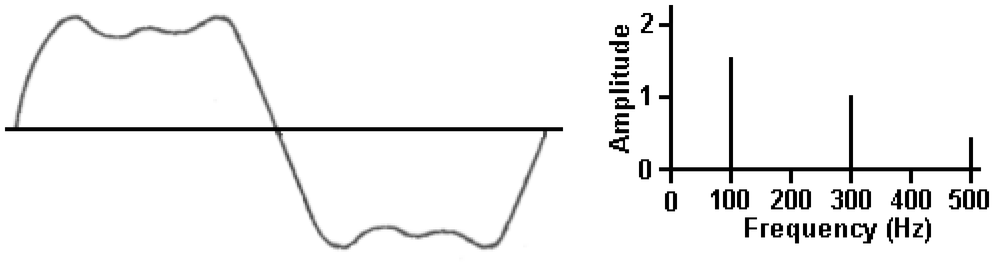
$$\{1, \cos t, \cos 2t, \ldots, \cos nt, \sin t, \sin 2t, \ldots, \sin nt\} \tag{5}$$

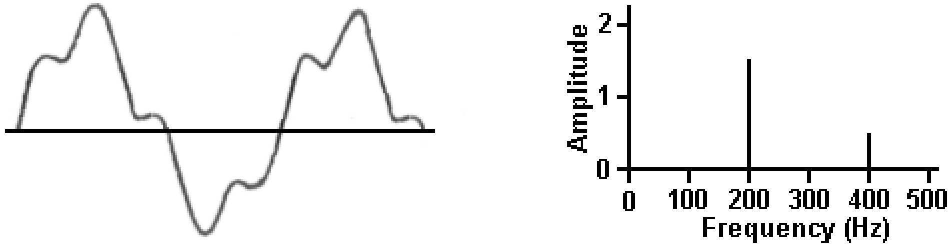
is orthogonal with respect to the inner product

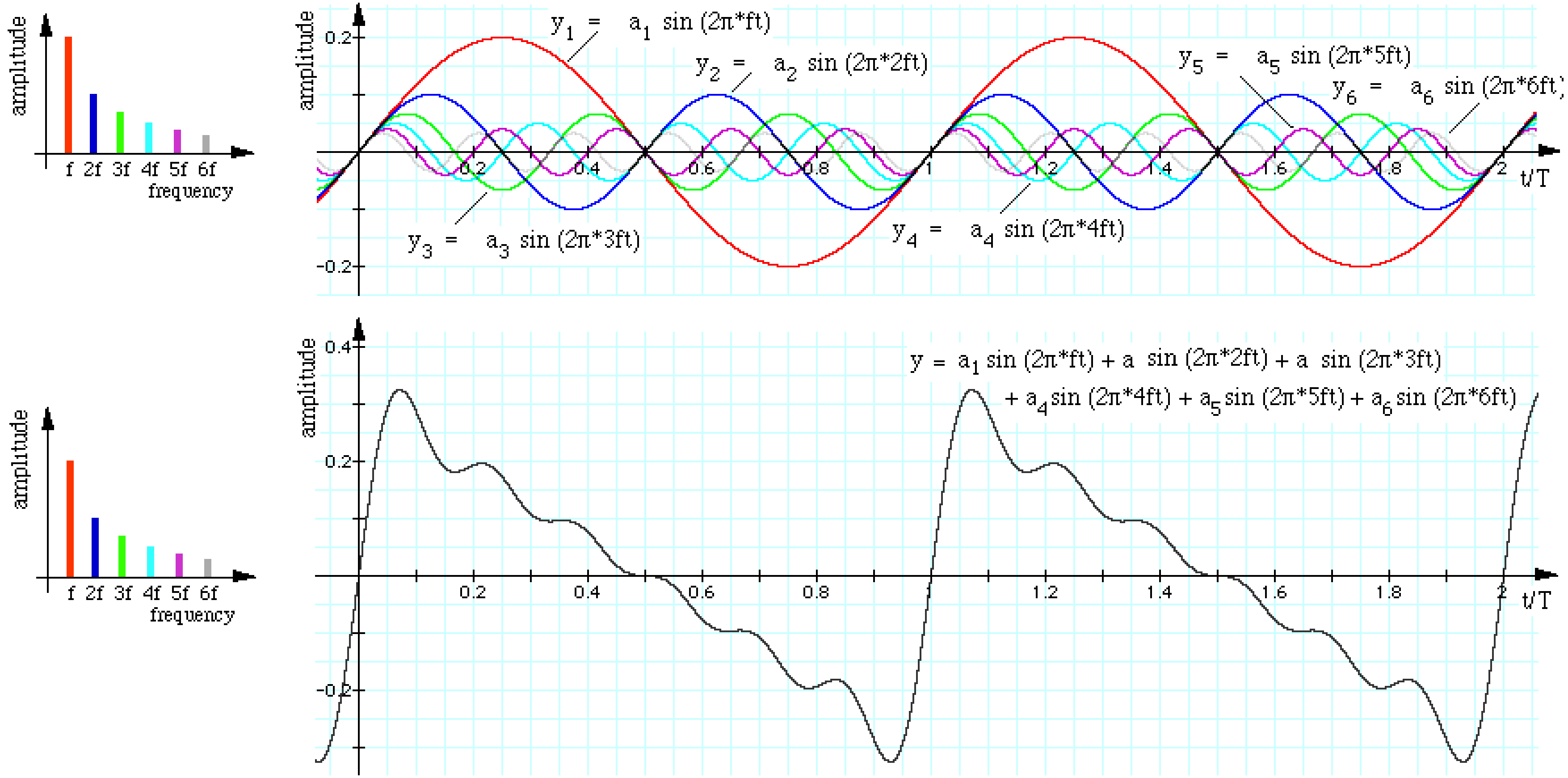
$$\langle f, g \rangle = \int_0^{2\pi} f(t)g(t)\,dt \tag{6}$$


Amplitude
0
1
2
0 100 200 300 400 500
Frequency (Hz)

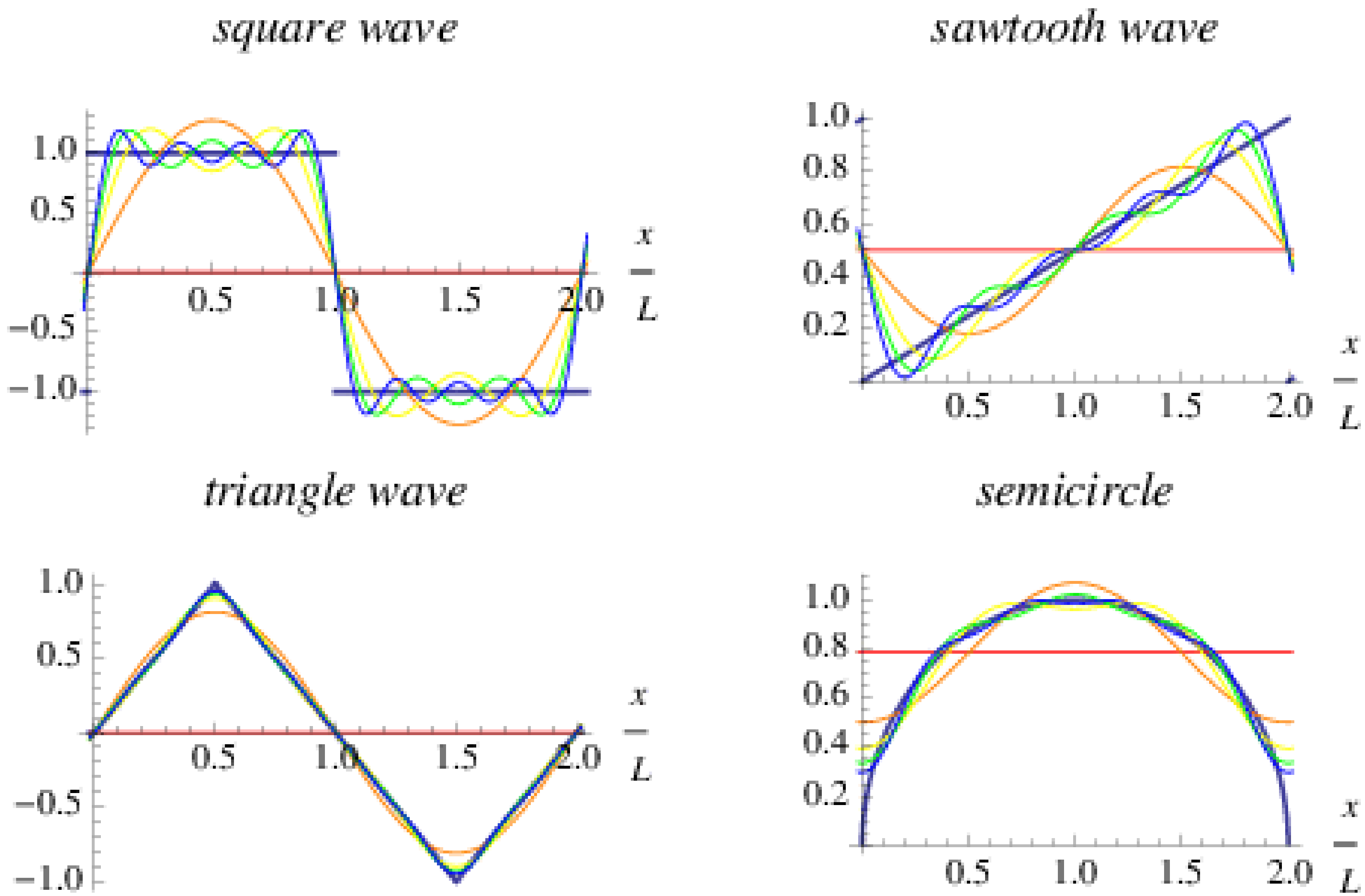

Amplitude
0
1
2
0 100 200 300 400 500
Frequency (Hz)

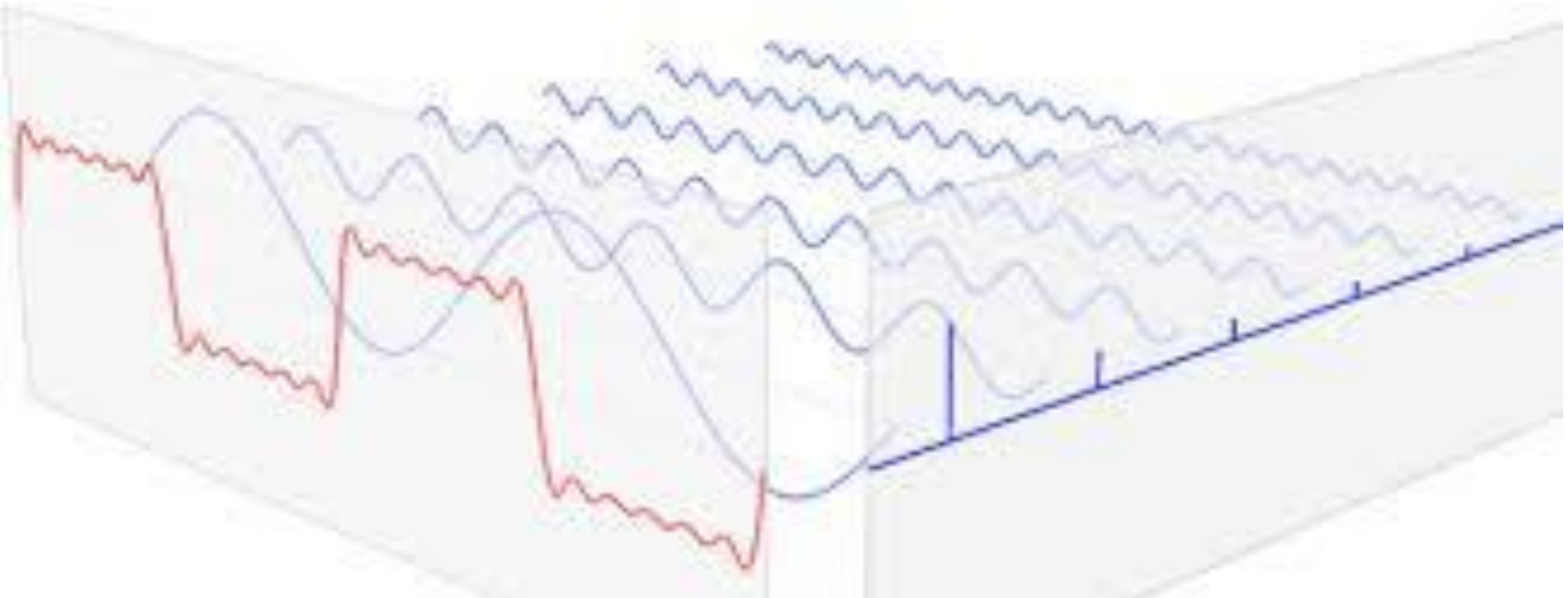

Amplitude
0
1
2
0 100 200 300 400 500
Frequency (Hz)

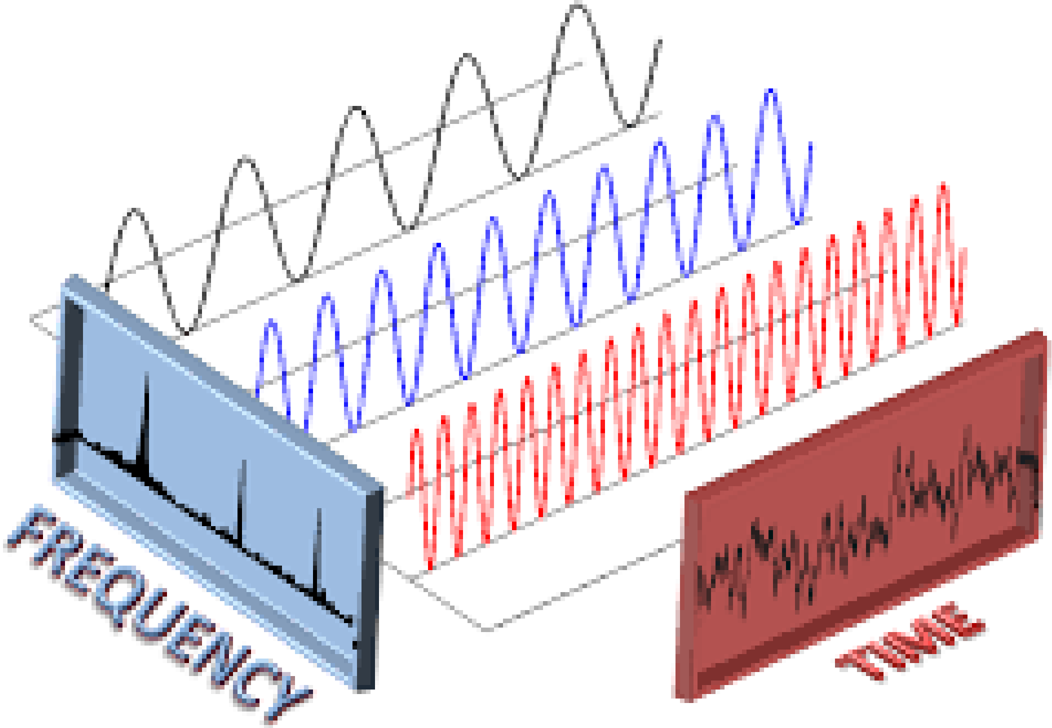

Amplitude
0
1
2
0 100 200 300 400 500
Frequency (Hz)

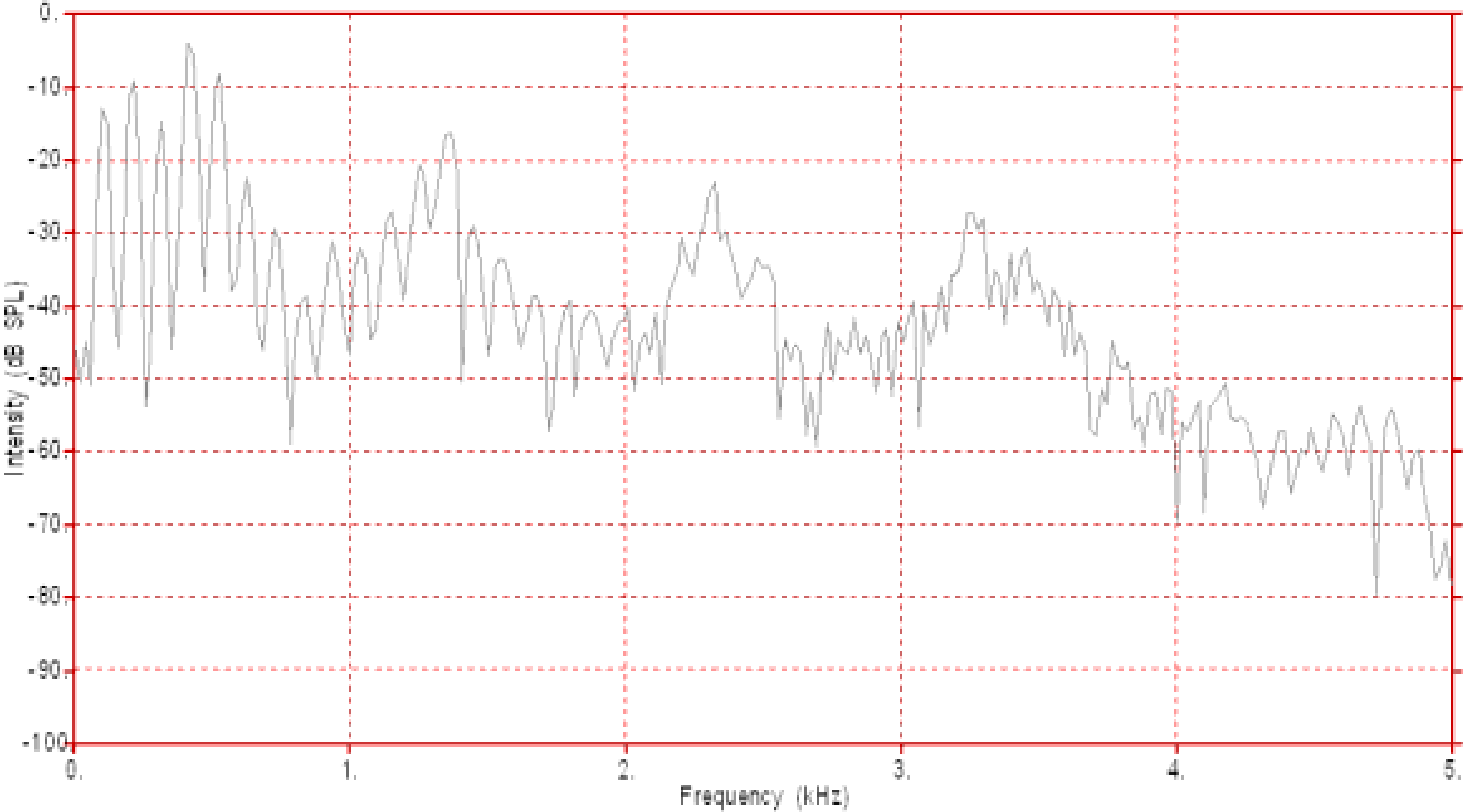


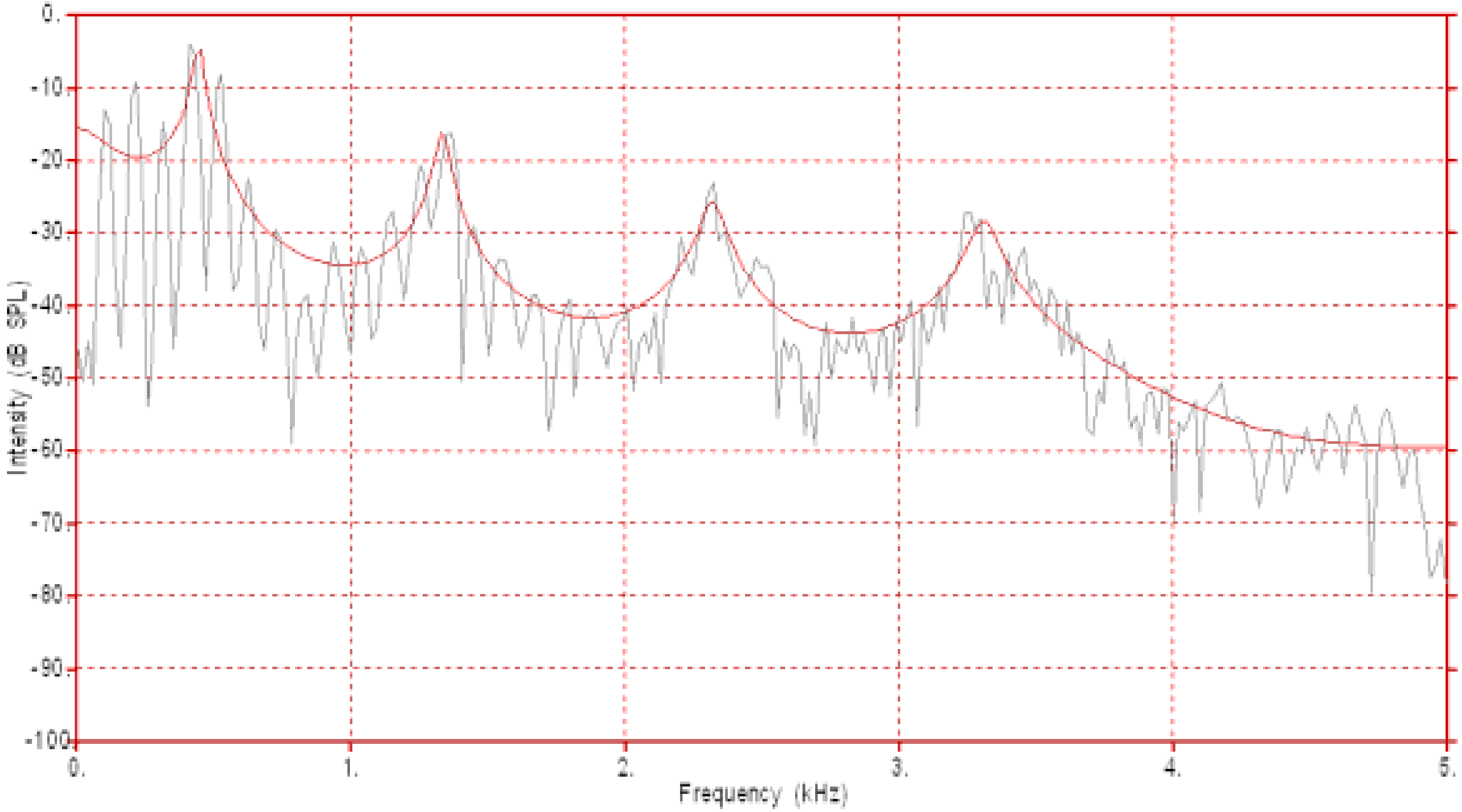
amplitude
f 2f 3f 4f 5f 6f
frequency
0.2
-0.2
$y_1 = a_1 \sin(2\pi * ft)$
$y_2 = a_2 \sin(2\pi * 2ft)$
$y_3 = a_3 \sin(2\pi * 3ft)$
$y_4 = a_4 \sin(2\pi * 4ft)$
$y_5 = a_5 \sin(2\pi * 5ft)$
$y_6 = a_6 \sin(2\pi * 6ft)$
0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2
t/T
0.4
$y = a_1 \sin(2\pi * ft) + a \sin(2\pi * 2ft) + a \sin(2\pi * 3ft)$
$+ a_4 \sin(2\pi * 4ft) + a_5 \sin(2\pi * 5ft) + a_6 \sin(2\pi * 6ft)$


square wave
sawtooth wave
triangle wave
semicircle
1.0
0.5
−0.5
−1.0
0.8
0.6
0.4
0.2
1.5
2.0
x
L

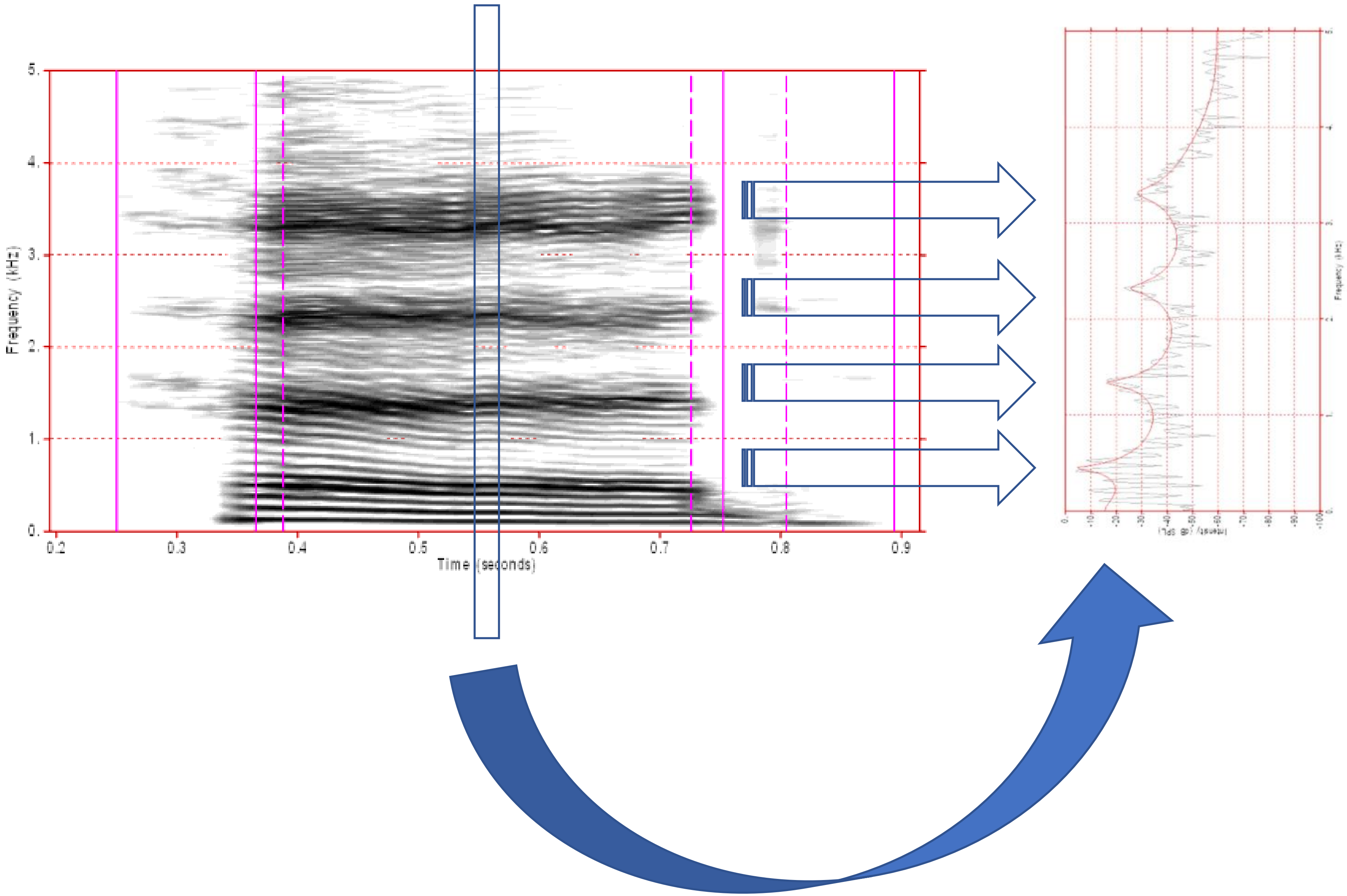



FREQUENCY
TIME


0.
-10.
-20.
-30.
-40.
-50.
-60.
-70.
-80.
-90.
-100.
Intensity (dB SPL)
0.
1.
2.
3.
4.
5.
Frequency (kHz)

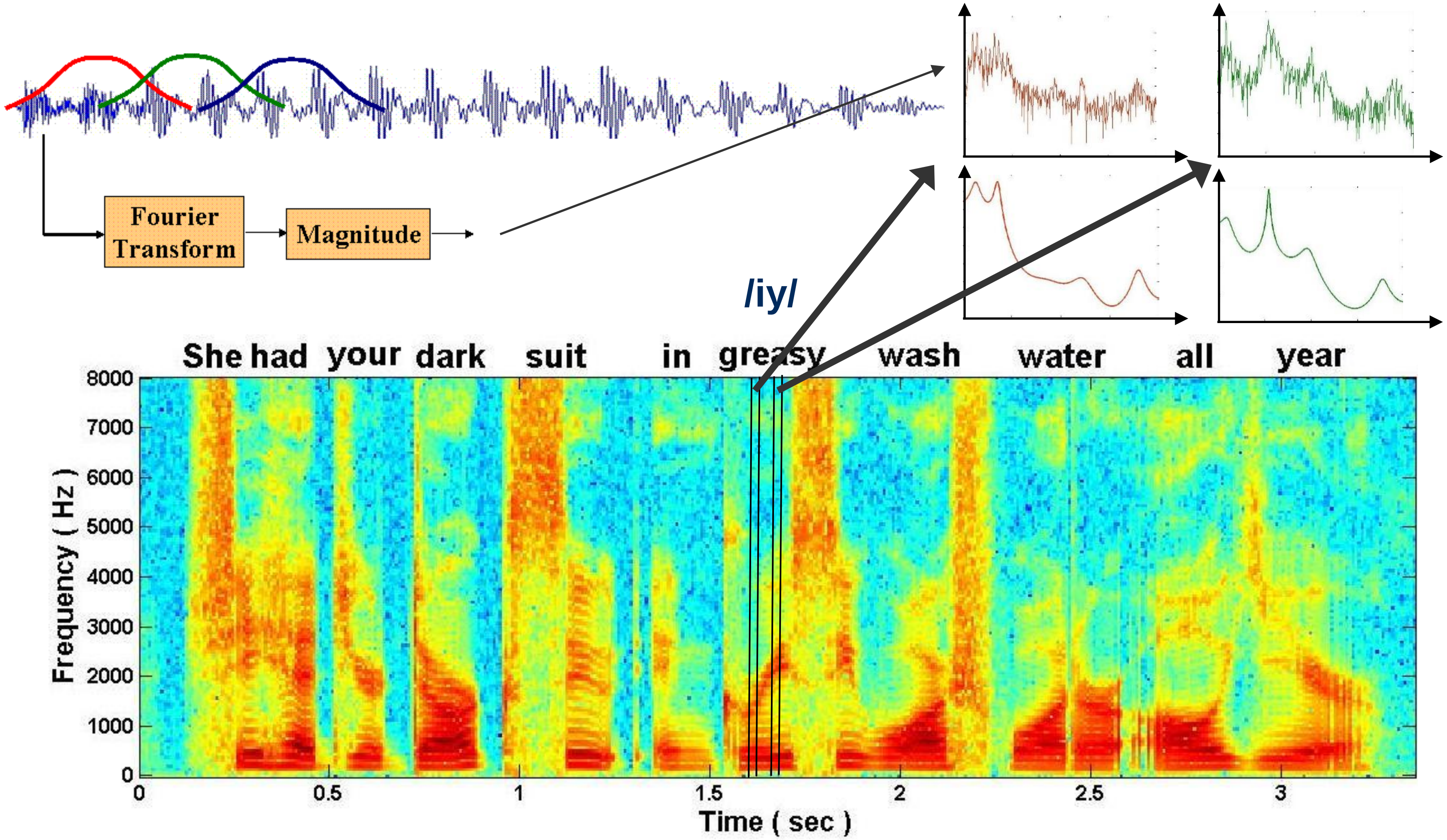


0.
-10.
-20.
-30.
-40.
-50.
-60.
-70.
-80.
-90.
-100.
Intensity (dB SPL)
0.
1.
2.
3.
4.
5.
Frequency (kHz)


Frequency (kHz)
Time (seconds)

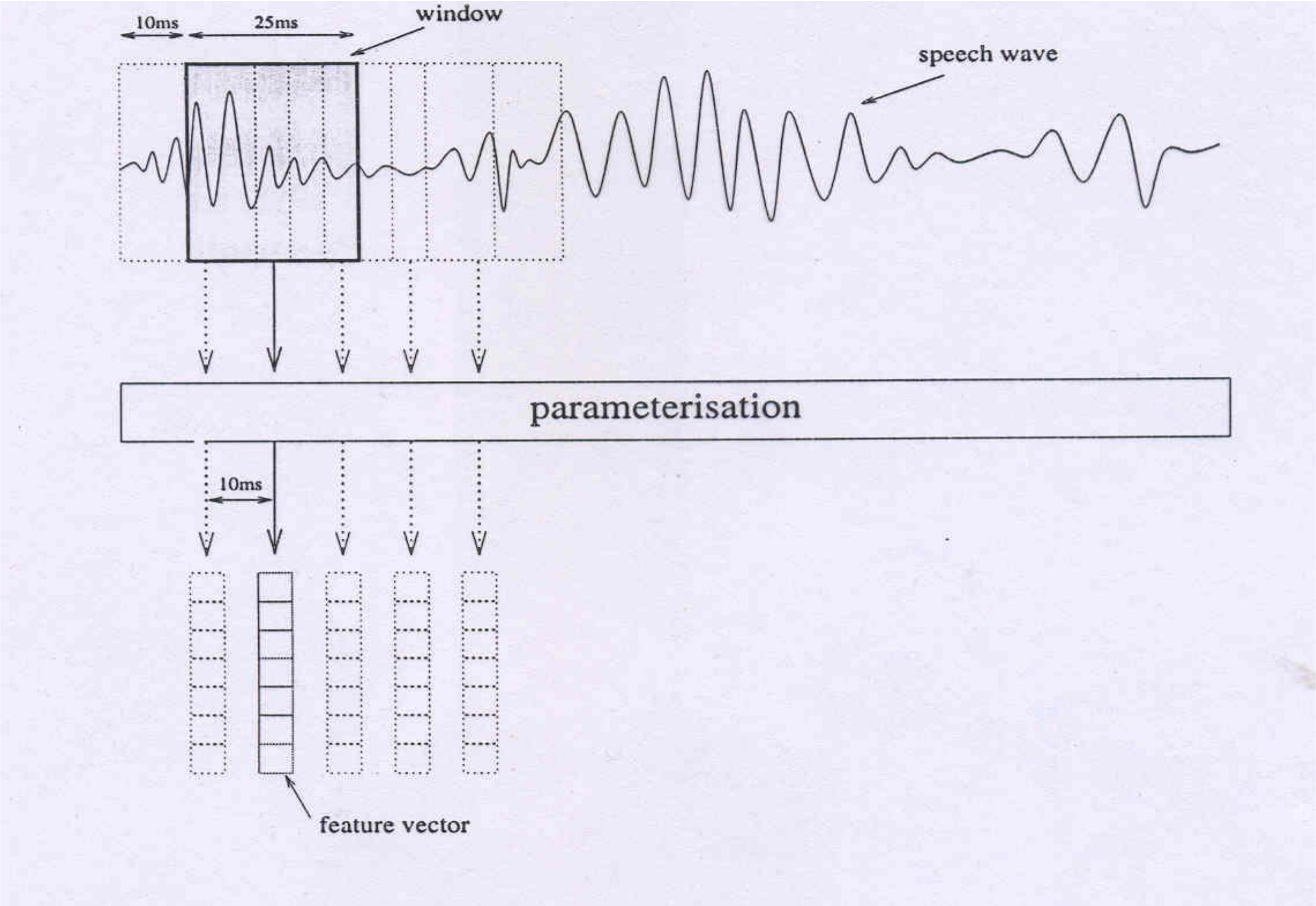
## Spectrogram

- Speech is a continuous evolution of the vocal tract
- Spectrogram shows time-frequency evolution
- Represented as a time-series of short-time spectra

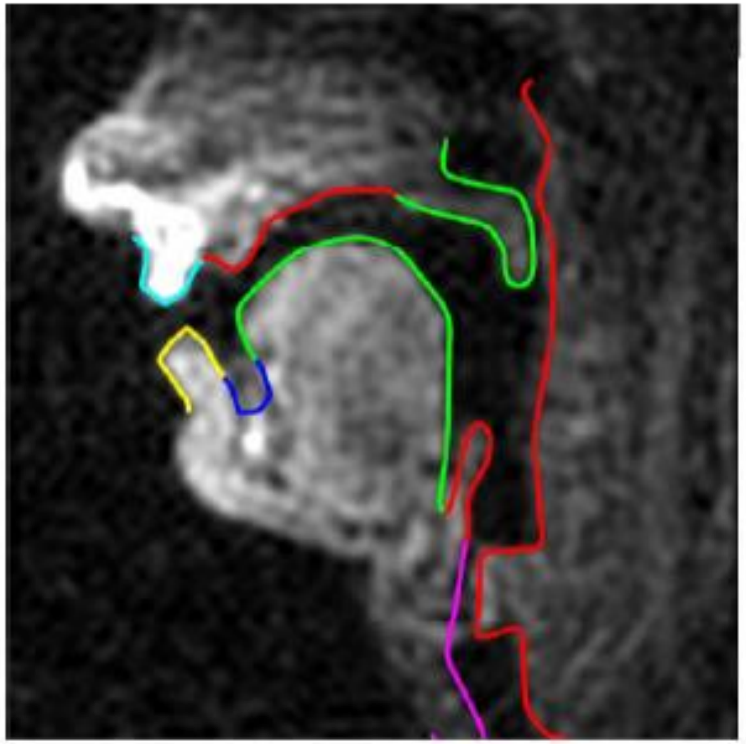


# Auto-regressive Model and Speech Spectrum

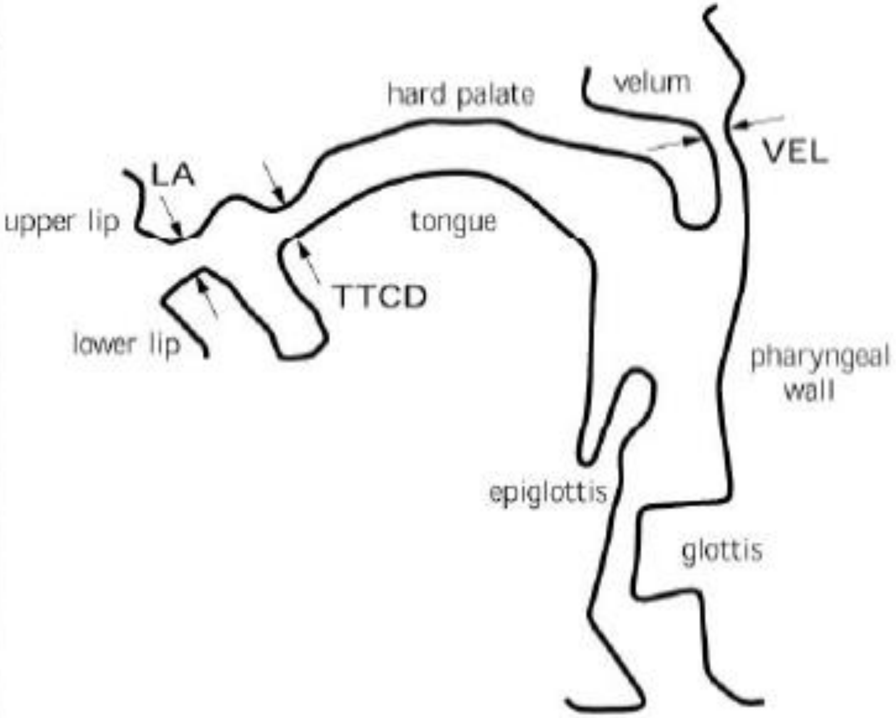
# Short-time Analysis and Parameterization



# MRI of Speech (Prof. Shri Narayanan, USC)



(a)

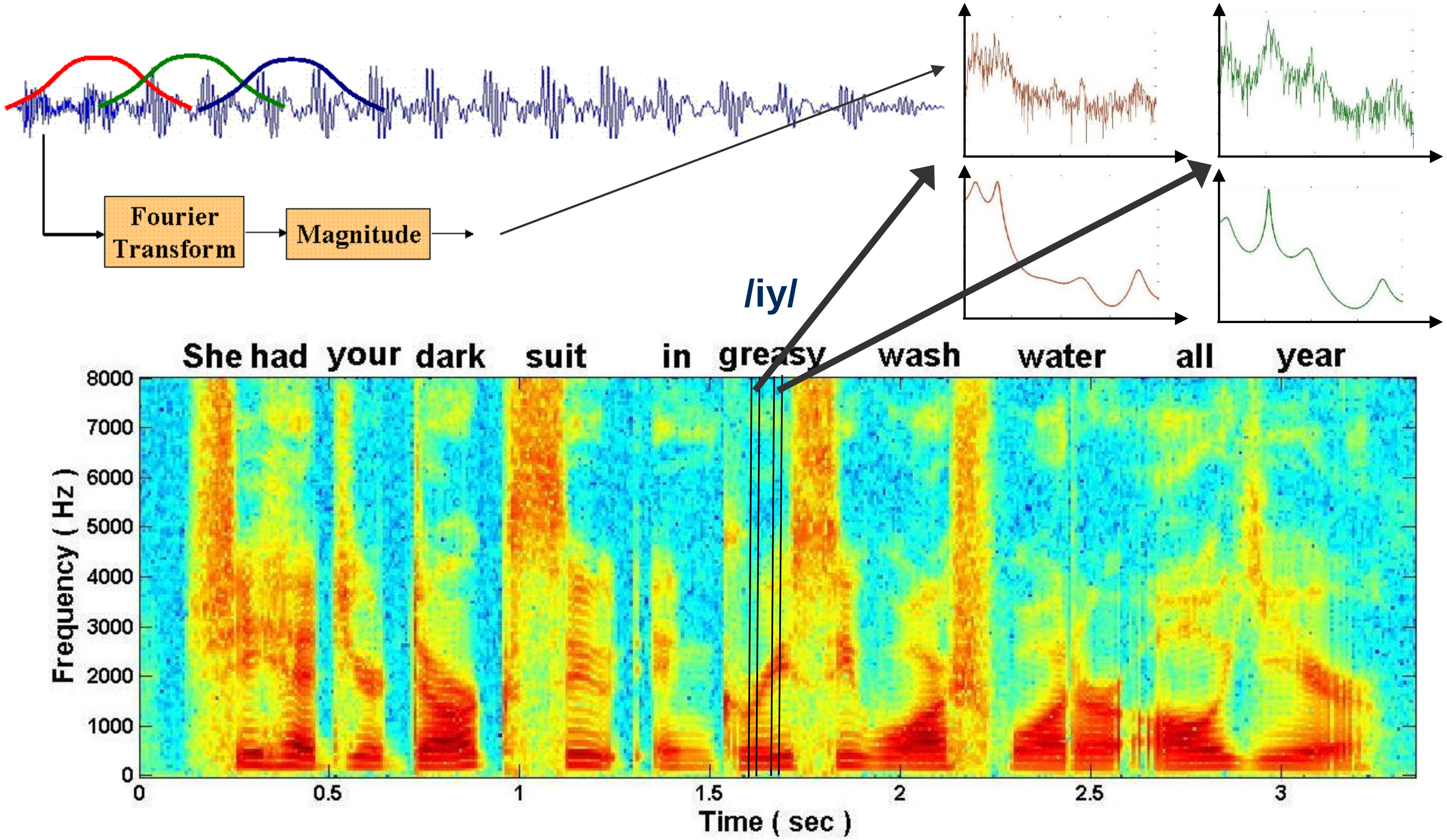



(b)


USC
SPAN

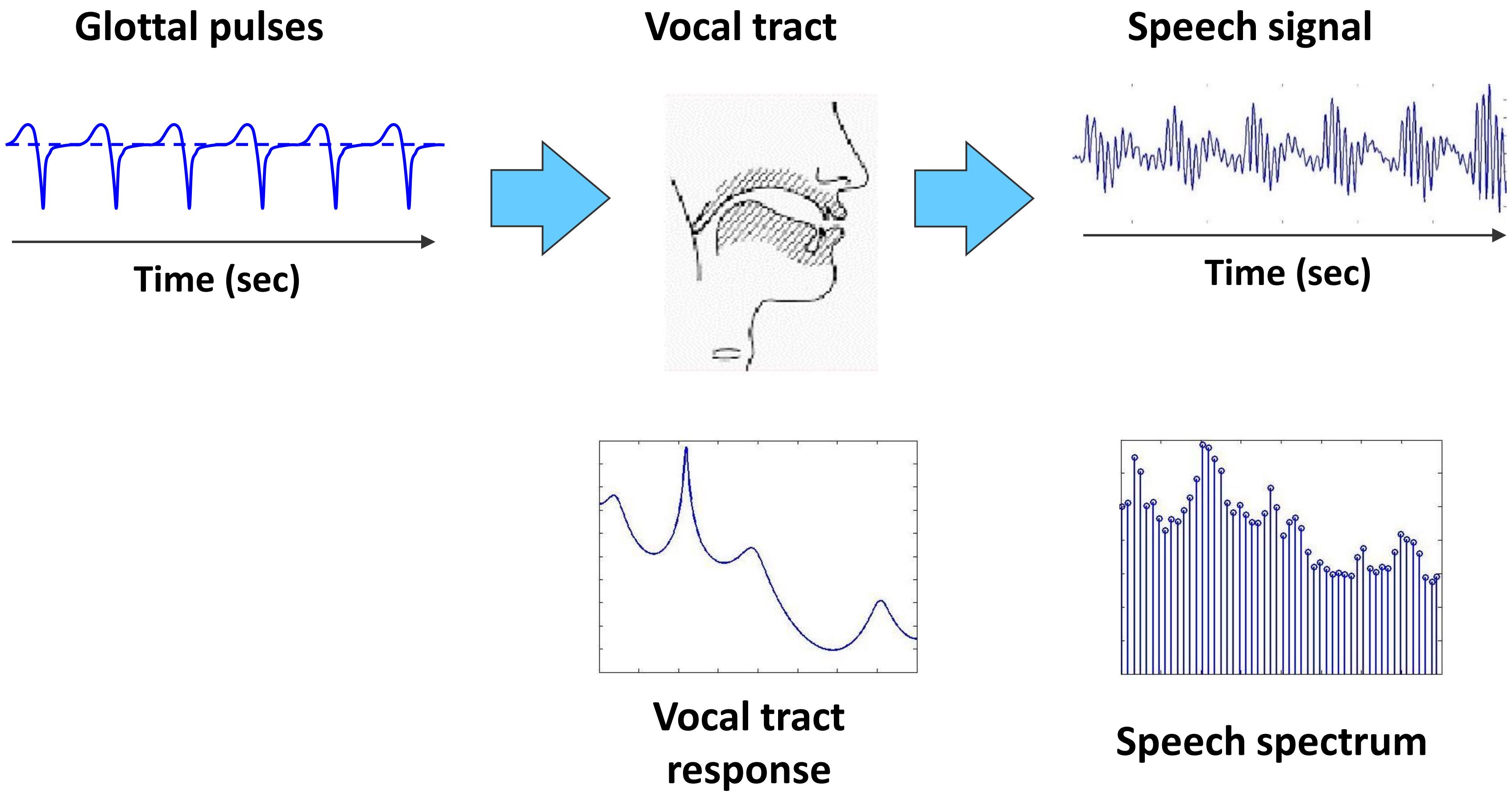
## Spectrogram

- Speech is a continuous evolution of the vocal tract
- Spectrogram shows time-frequency evolution
- Represented as a time-series of short-time spectra



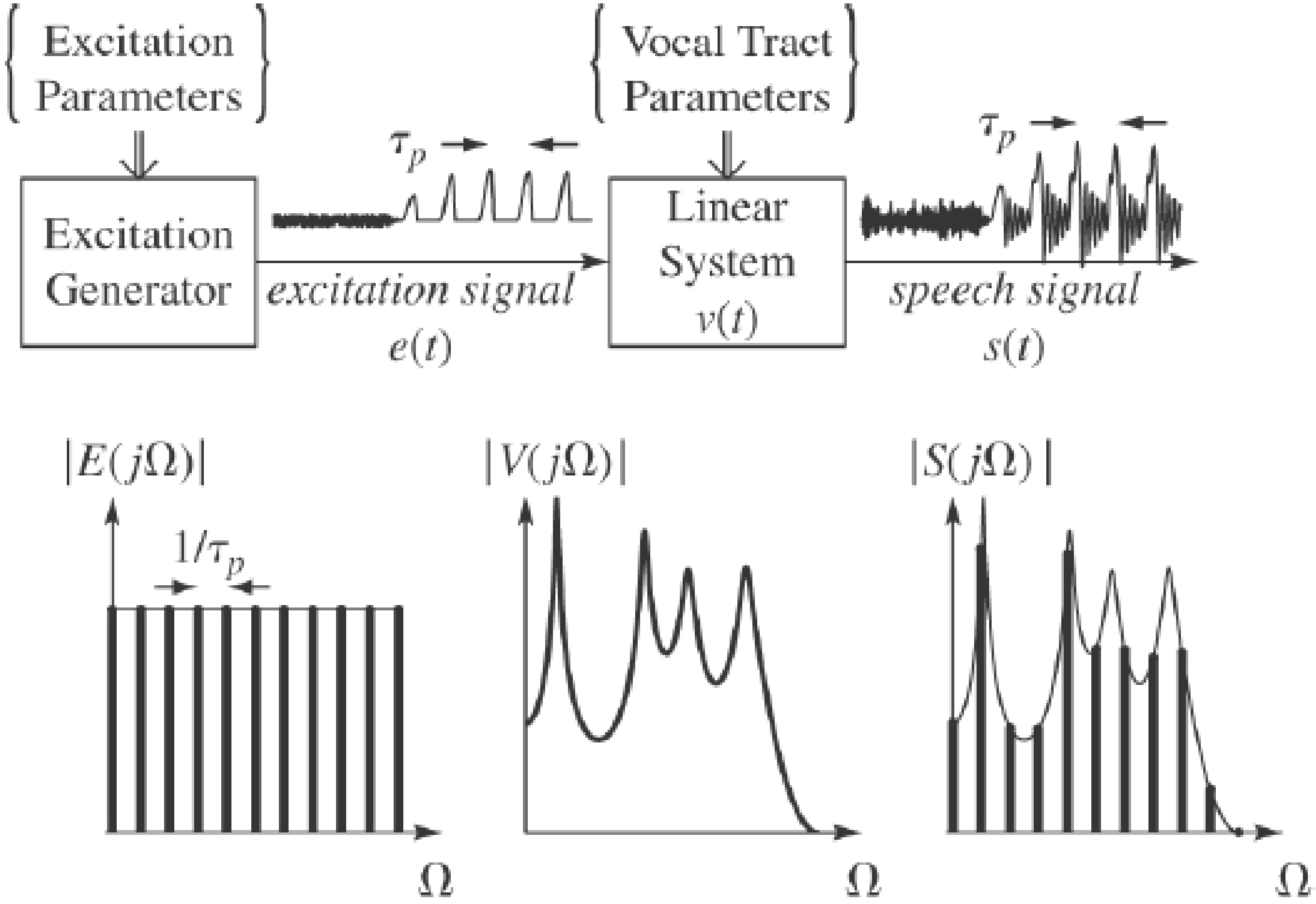
## Source-Filter Model

- Features based on speech production model: Source-filter interaction
  - – Anatomical structure (vocal tract / glottis) conveyed in speech spectrum

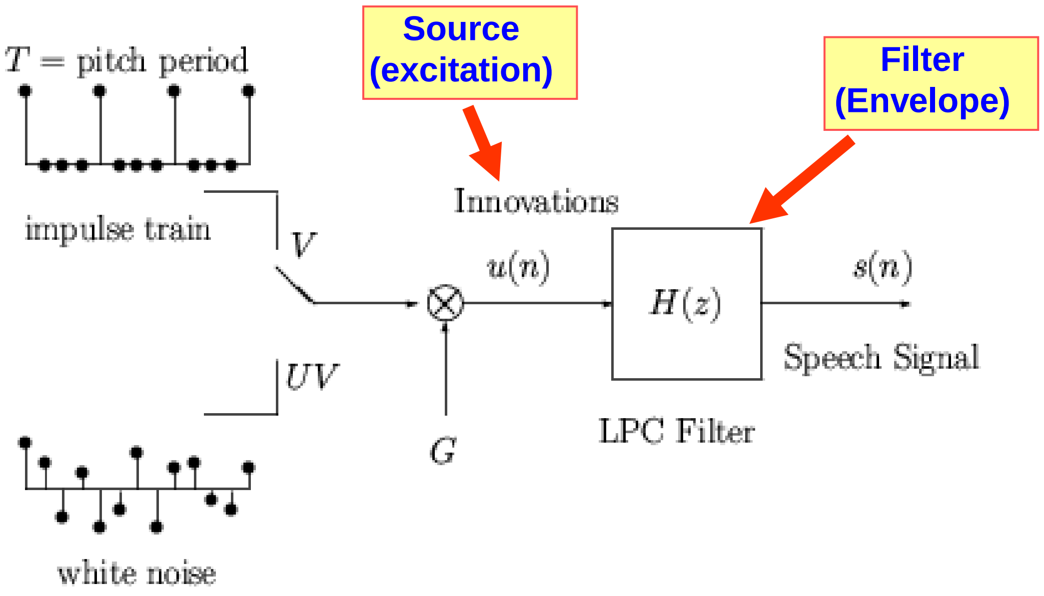



To Quatieri and Rab – Slides ➔

# Source-System Model of Speech Production



# Linear Prediction based Speech Production Model

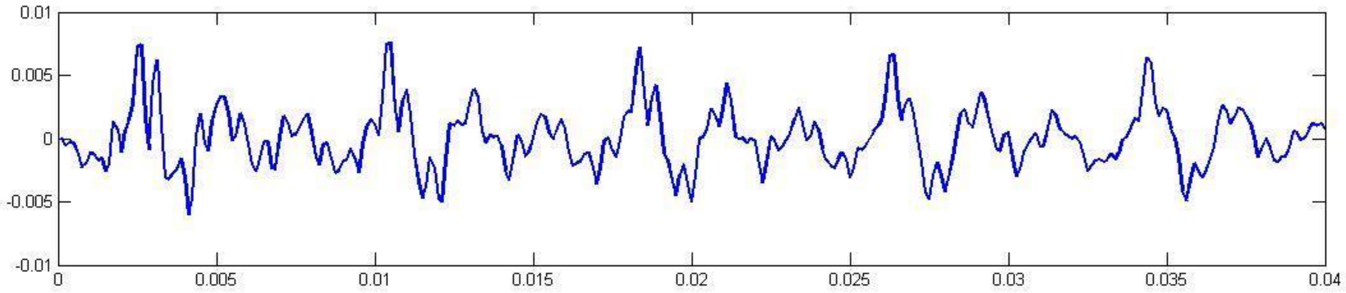



$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

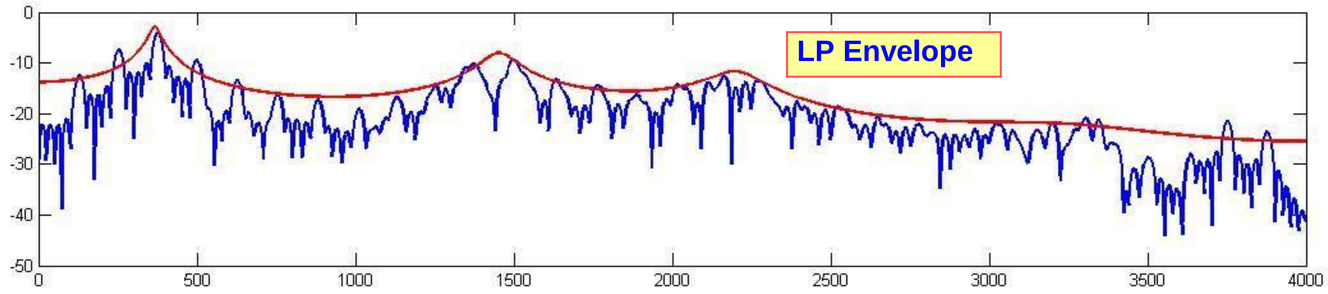
| | | |
|---|---|---|
| •Vocal Tract | ⟺ | • H(z) (LPC Filter) |
| •Air | ⟺ | • u(n) (Innovations) |
| •Vocal Cord Vibration | ⟺ | • V (voiced) |
| •Vocal Cord Vibration Period | ⟺ | • T (pitch period) |
| •Fricatives and Plosives | ⟺ | • UV (unvoiced) |
| •Air Volume | ⟺ | • G (gain) |

# LP Analysis: Envelope (Filter) & Excitation (Source)

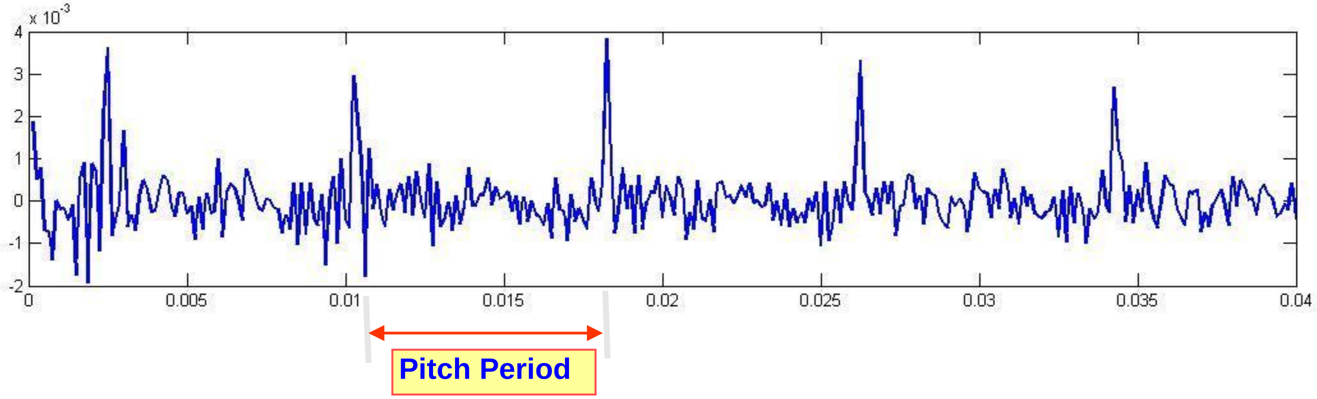
**Speech Signal S(n)**



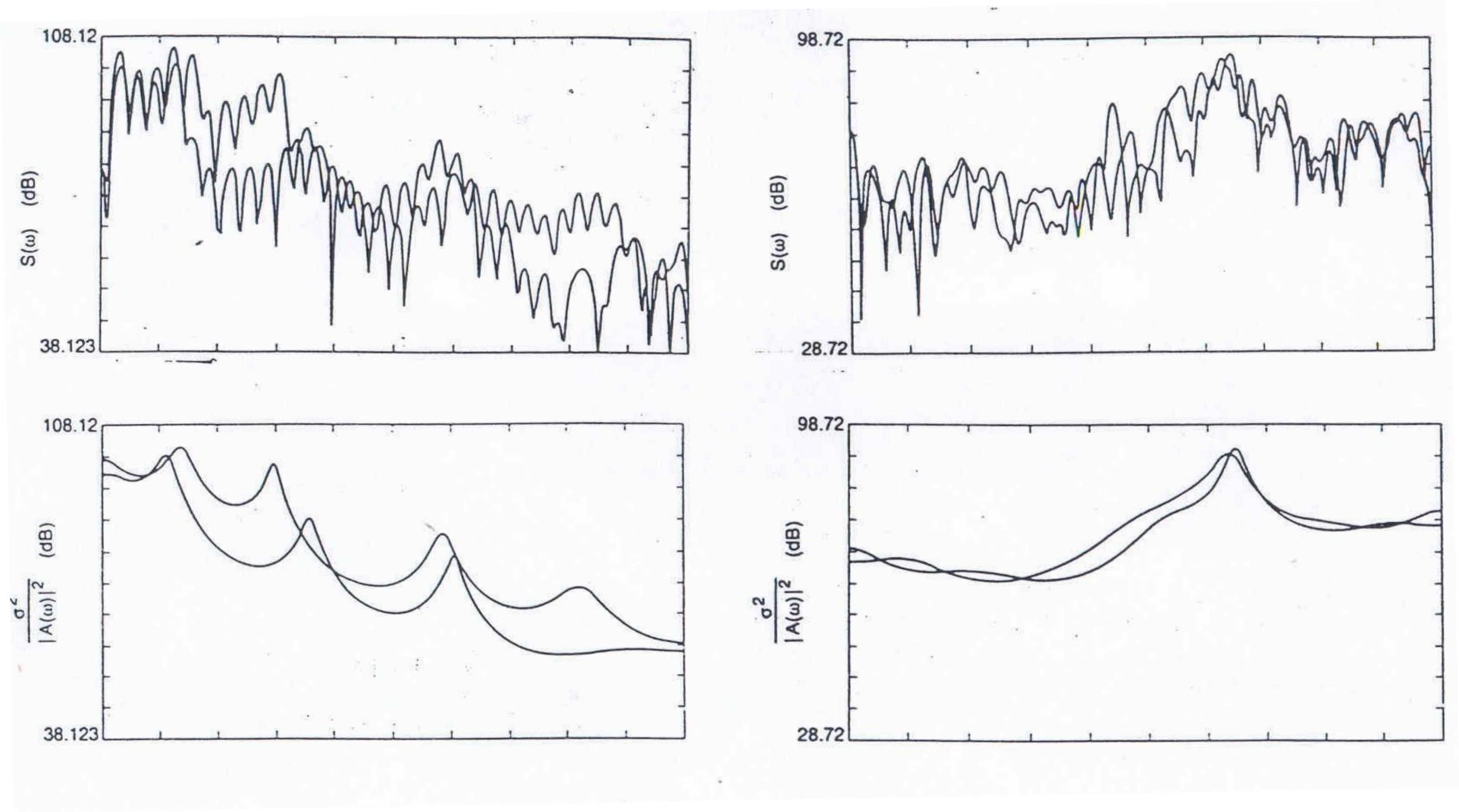

**S(w) with LP spectral envelope superimposed**



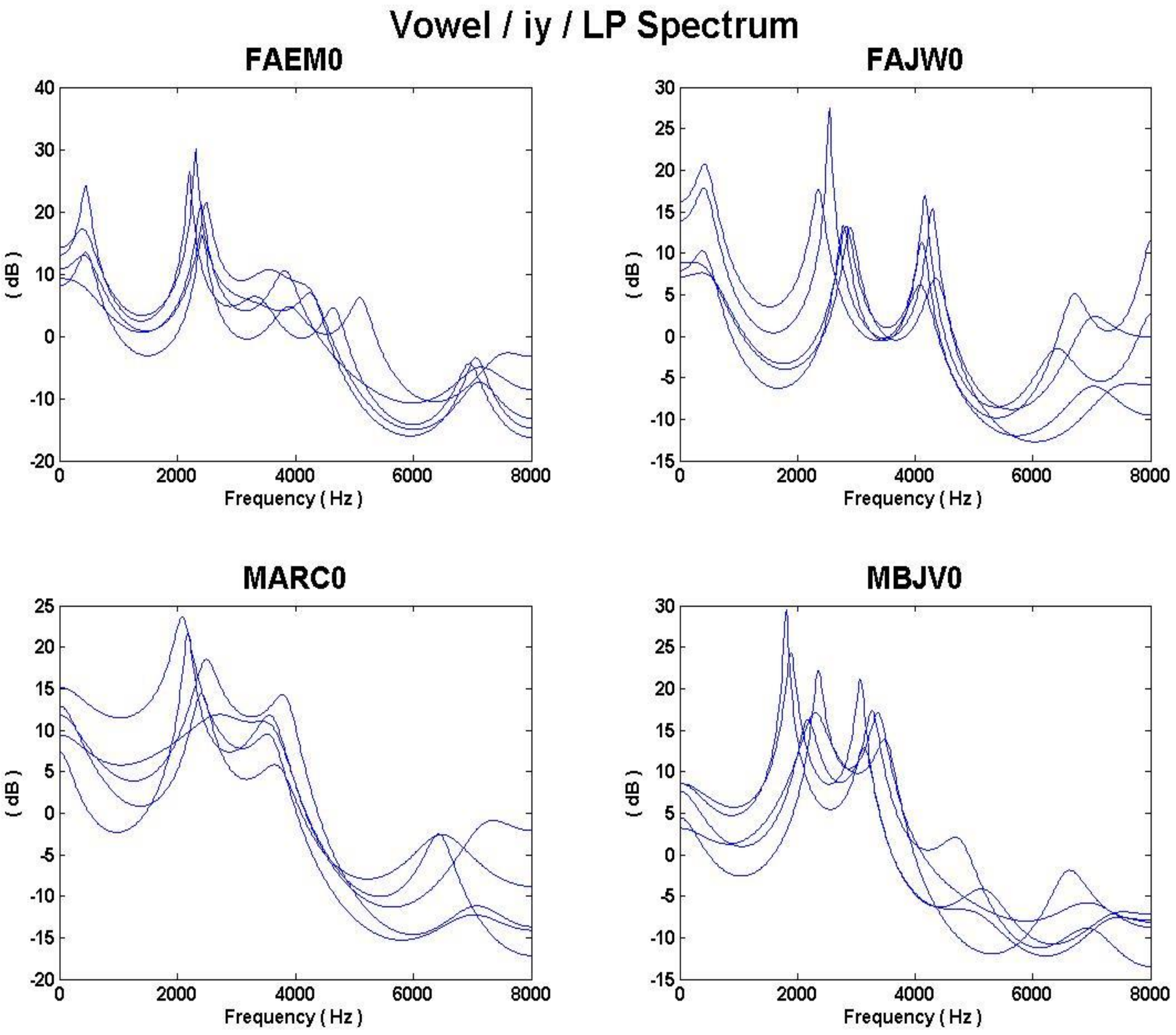

**Excitation Signal E(n)**



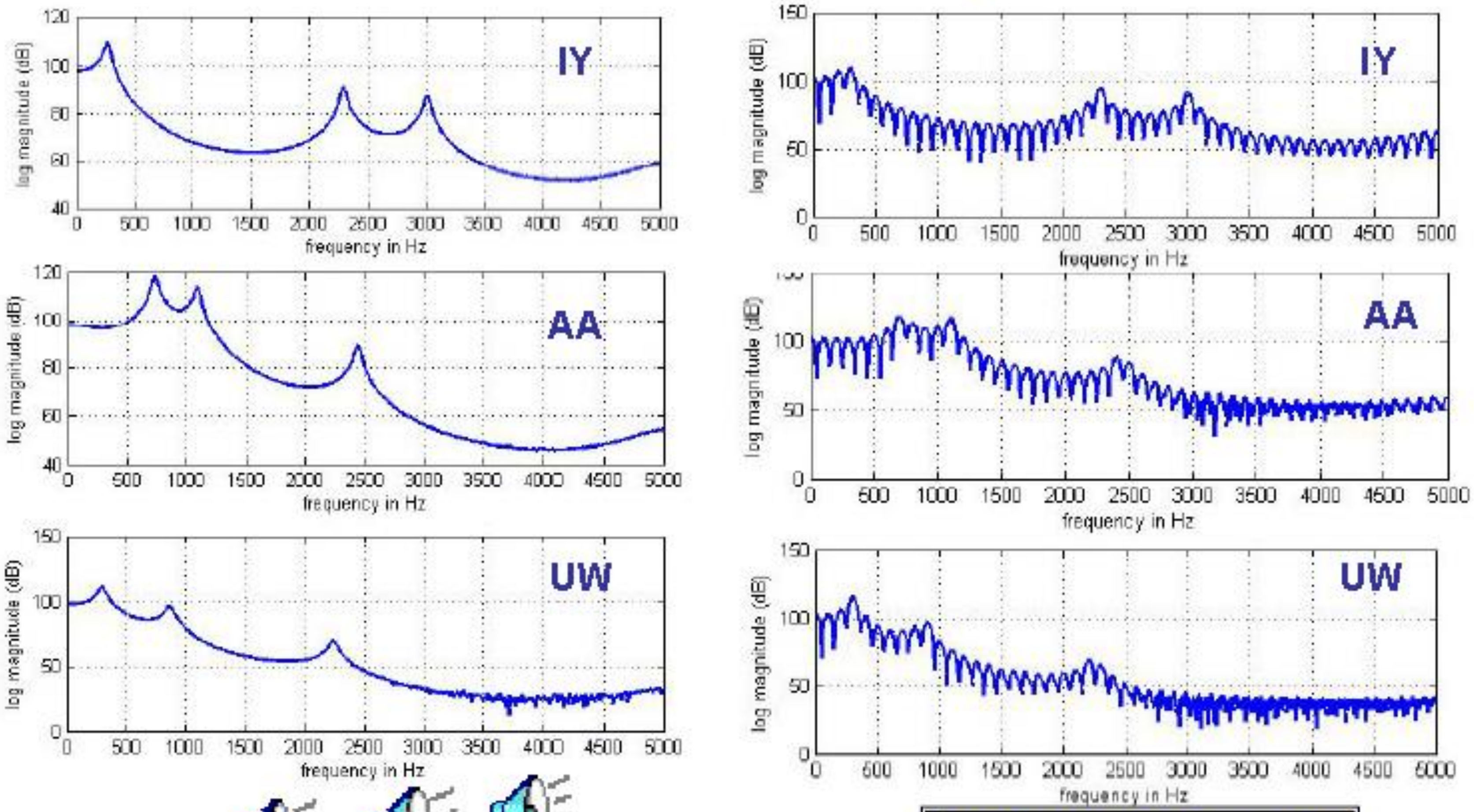
# Spectral slices



# Spectral Envelopes



# Canonic Vowel Spectra

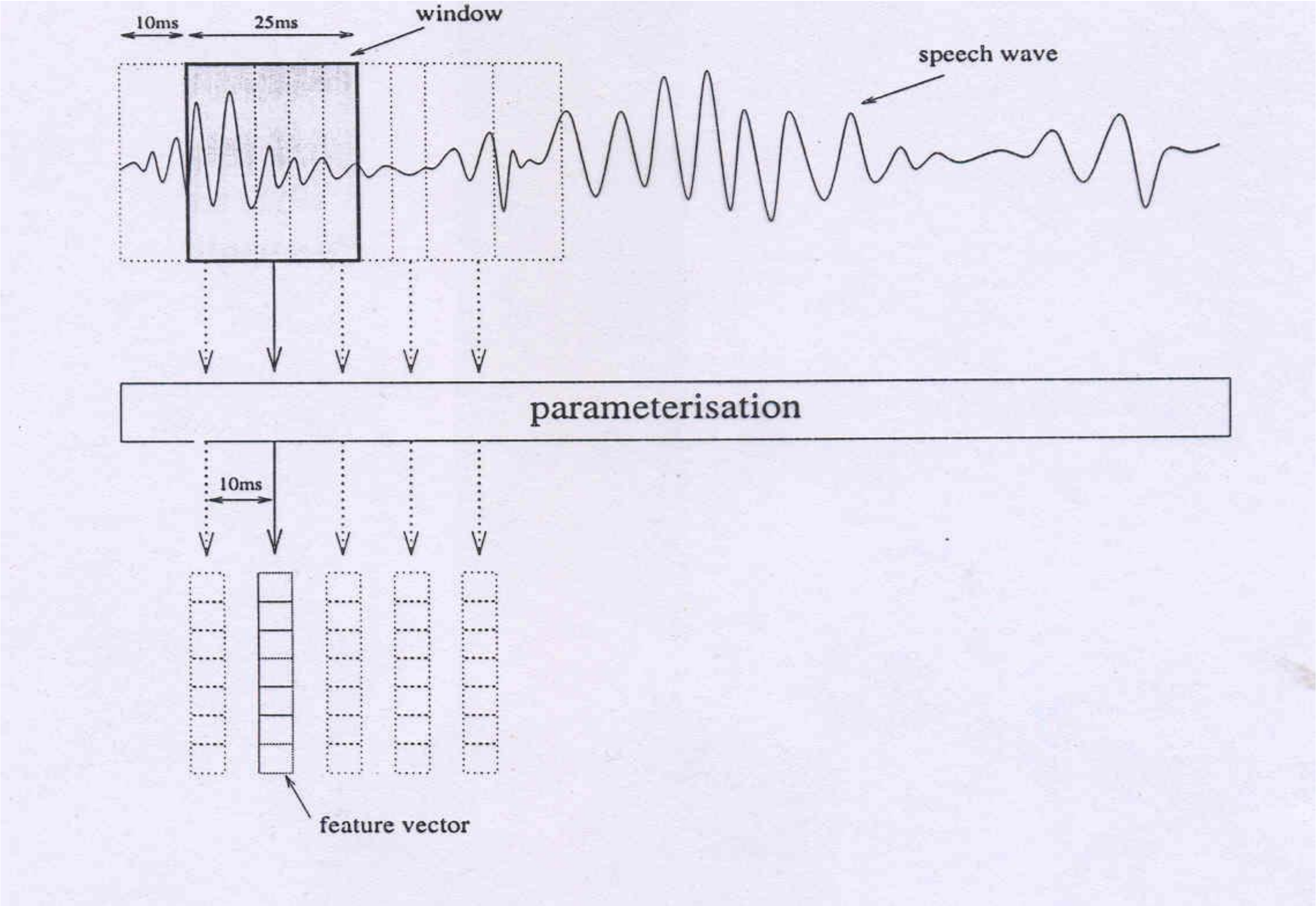




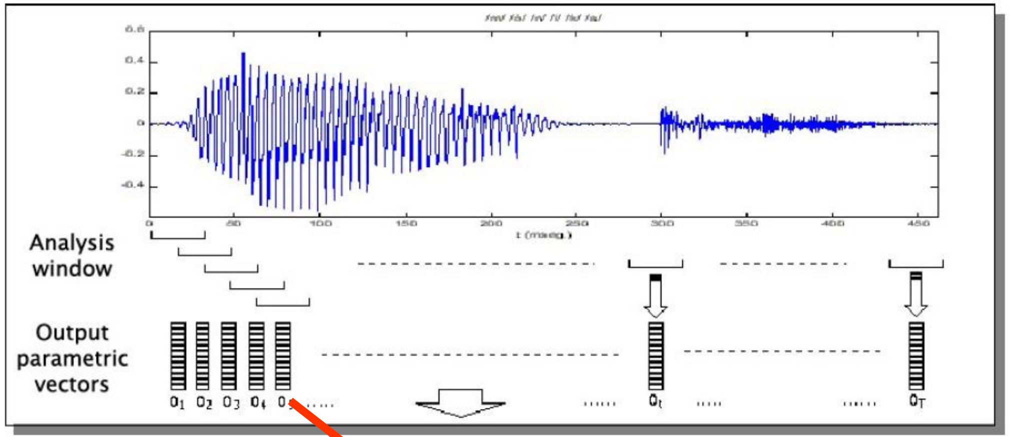
**100 Hz Fundamental**

54

# Short-time Analysis and Parameterization

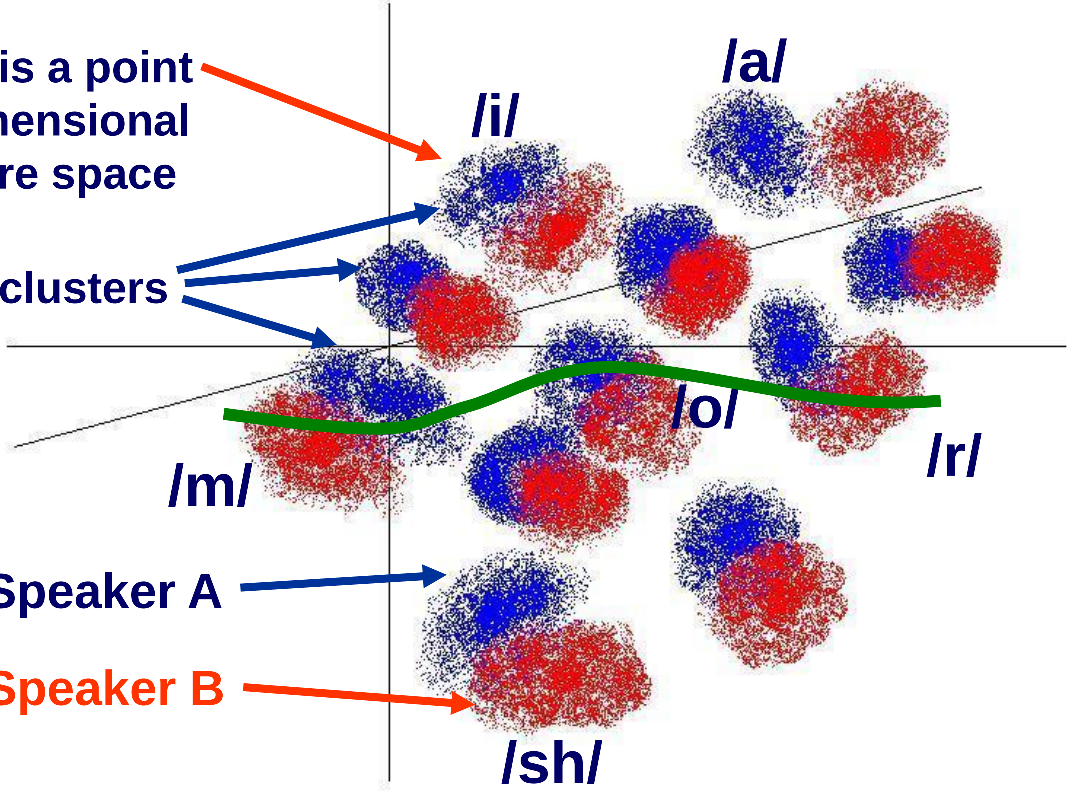


# Feature Space




Bag of vectors representation of speaker's acoustic space

**Each vector is a point in the 13-dimensional MFCC feature space**

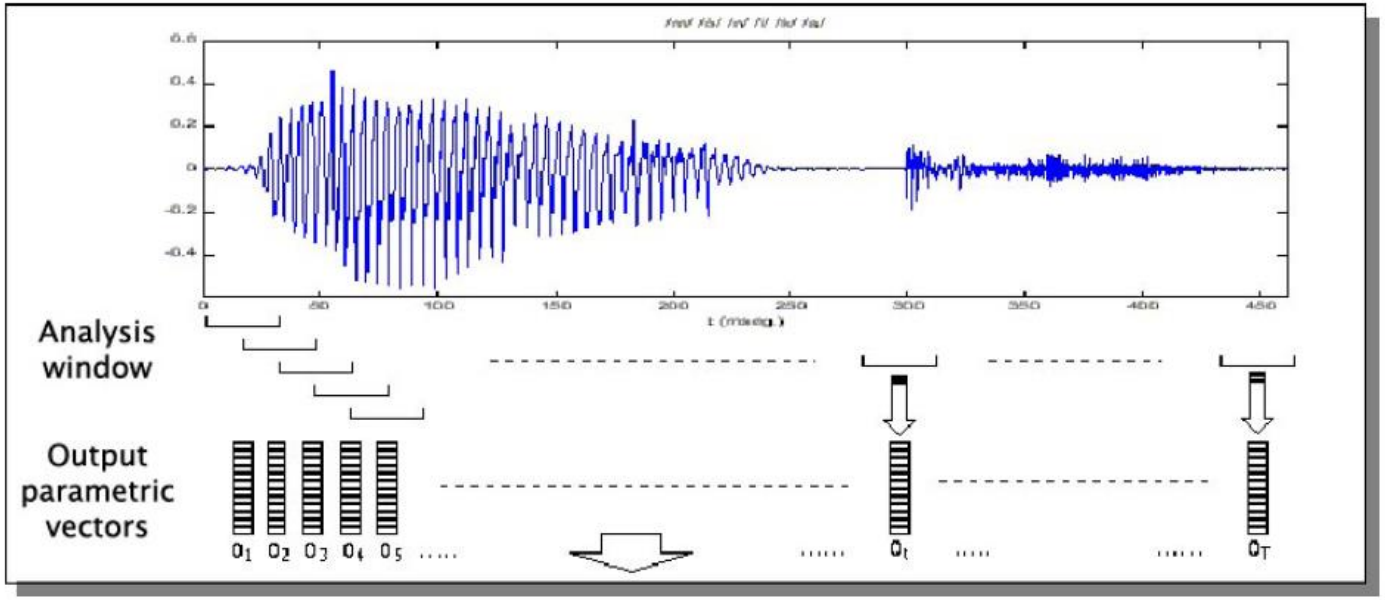



**Phone clusters**

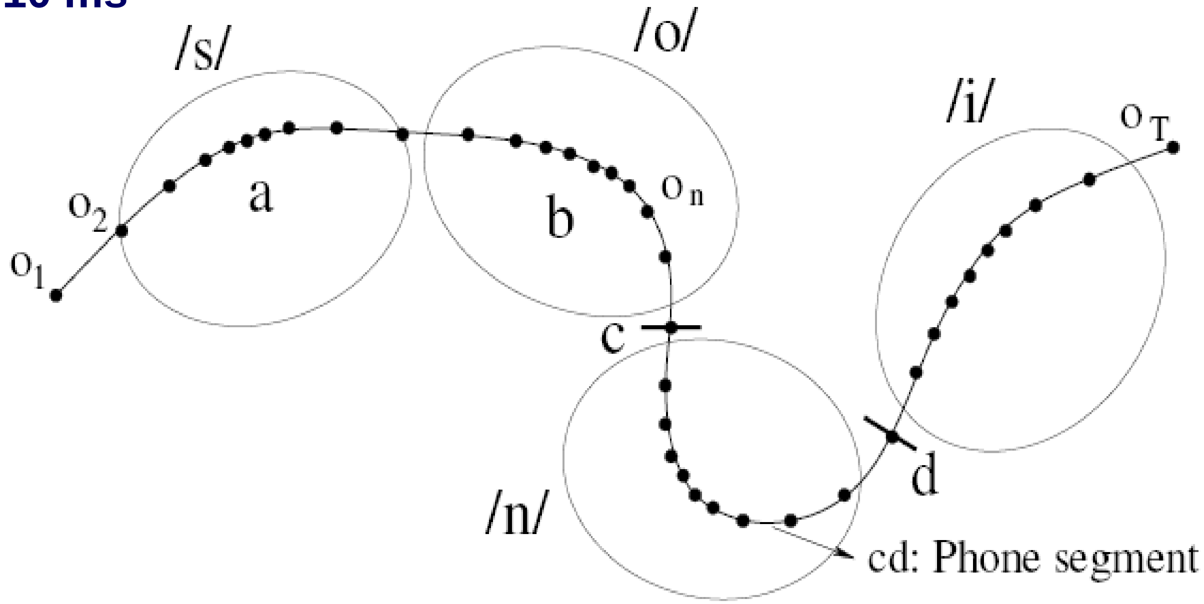
**Speaker A**

**Speaker B**

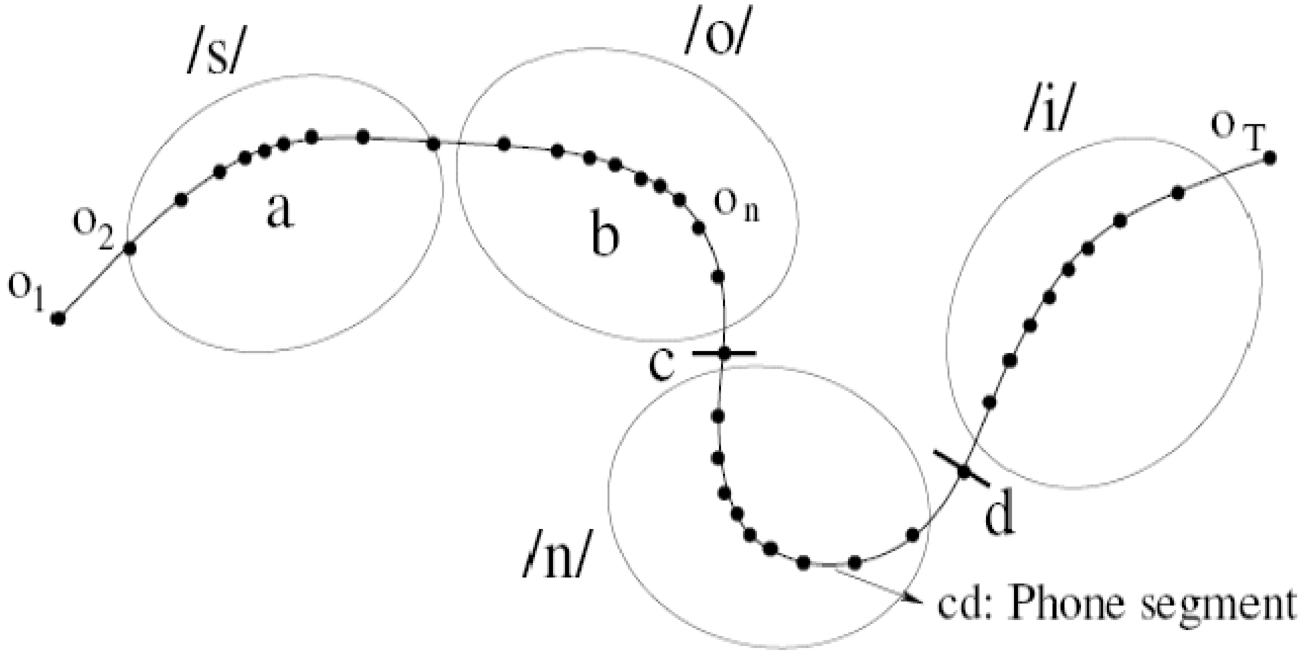
## Feature Space




One feature vector every 10 ms



# Feature Space

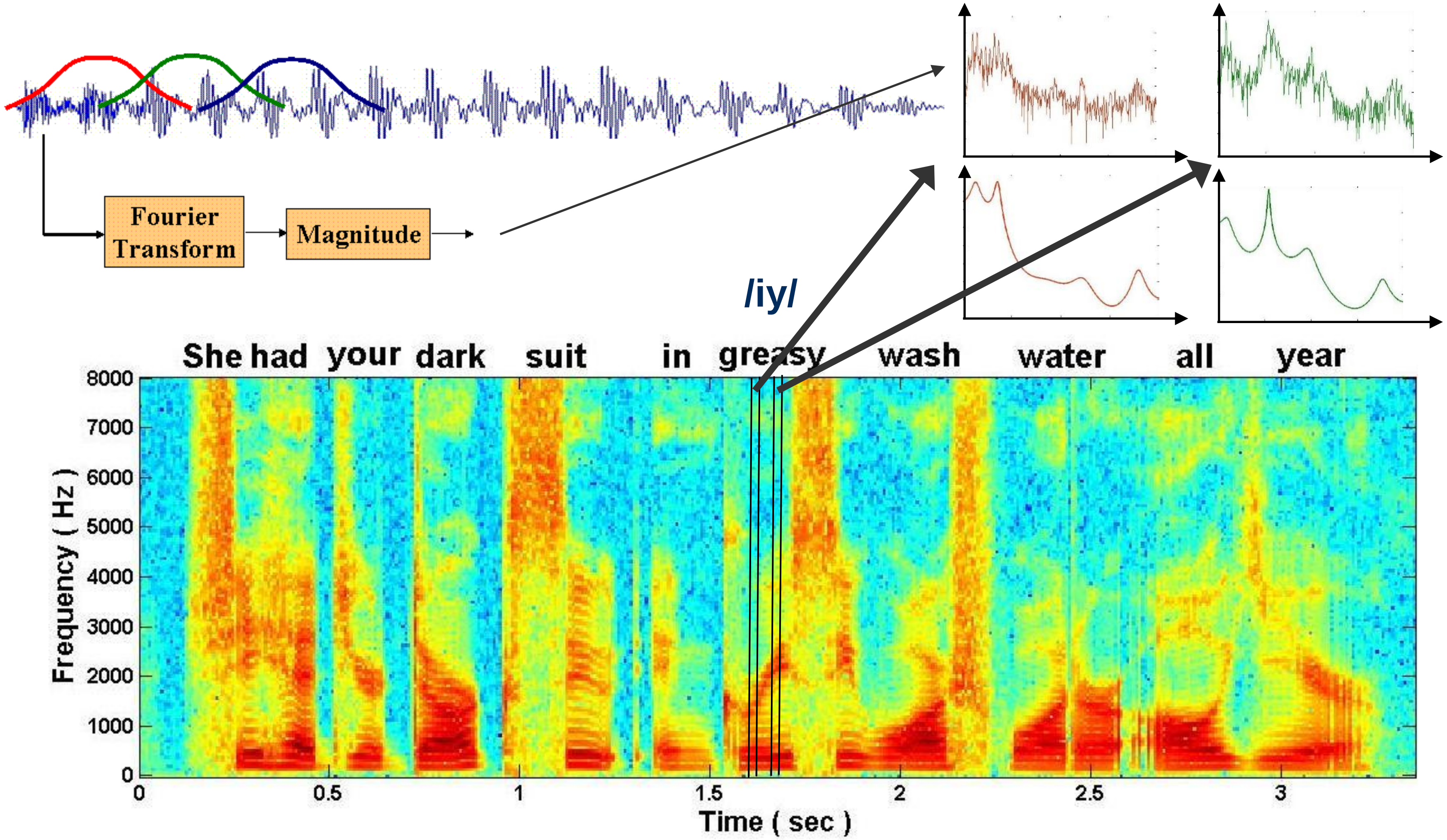



**SPEECH RECOGNITION ALGORITHMS**

- TAKE THIS FEATURE VECTOR SEQUENCE
- AS INPUT AND DETERMINE "WHAT HAS BEEN SAID"
- e.g. SEQUENCE OF PHONES / SEQUENCE OF WORDS etc.

## Spectrogram

- Speech is a continuous evolution of the vocal tract
- Spectrogram shows time-frequency evolution
- Represented as a time-series of short-time spectra



Thank you !!