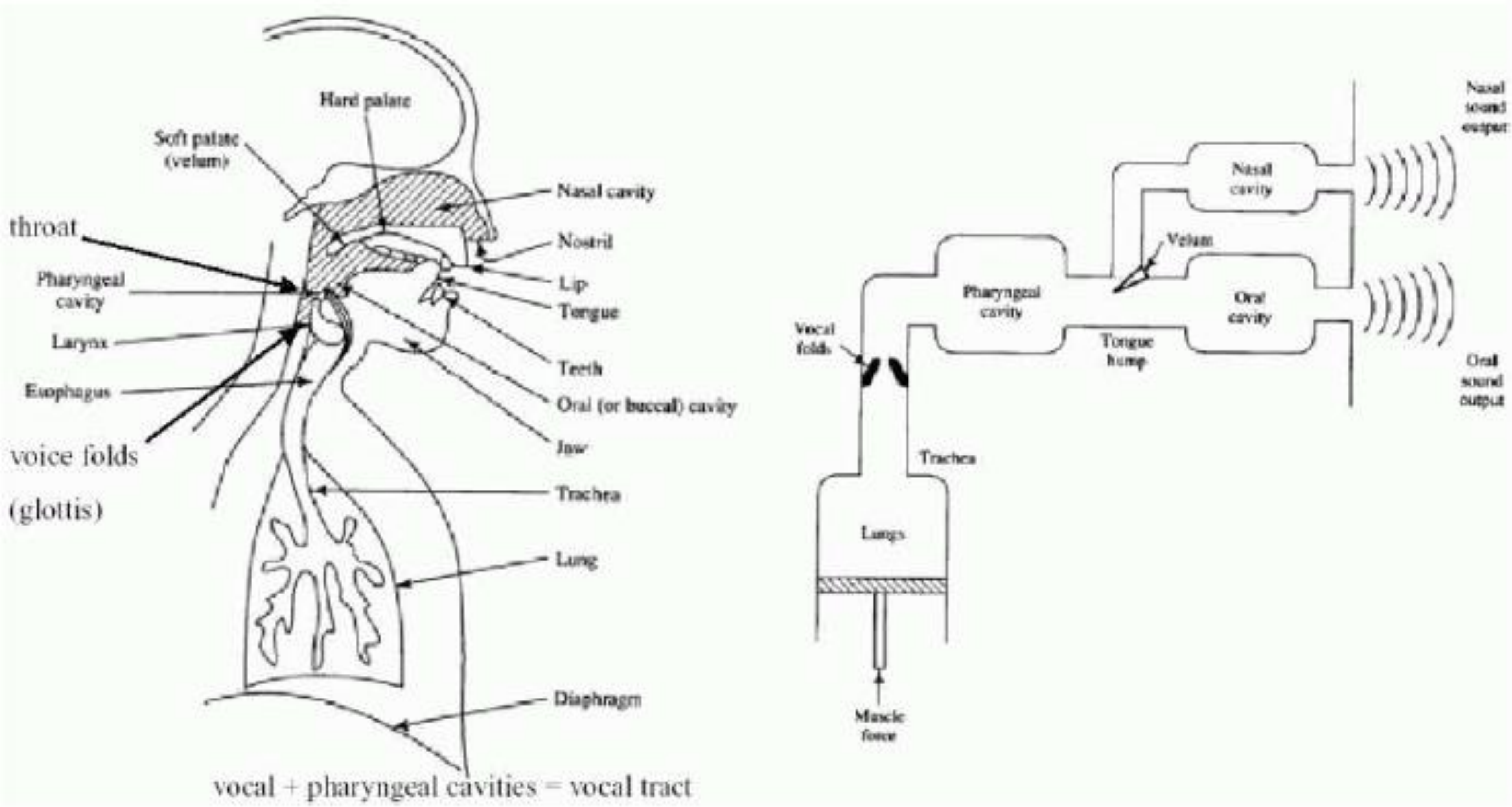
Speech Modeling, Analysis, Synthesis & Compression

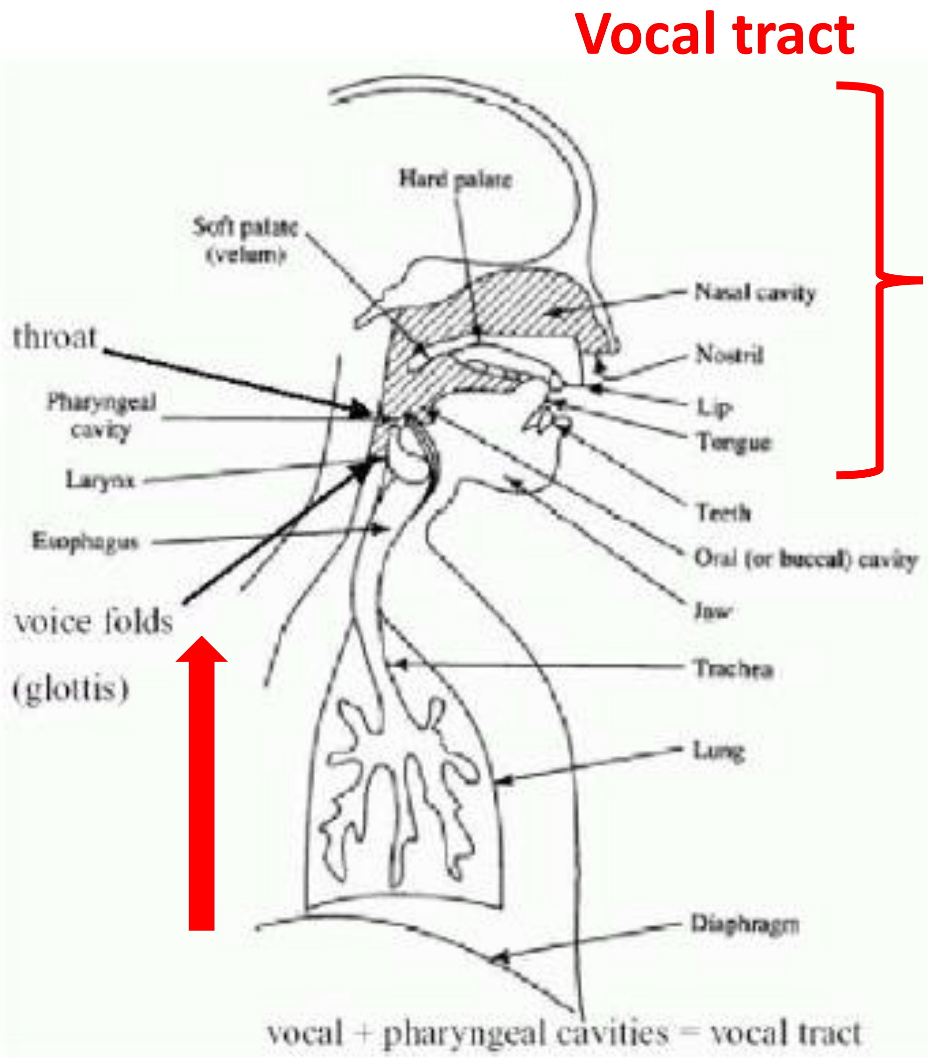
Tutorial for SSY150: Lab 1

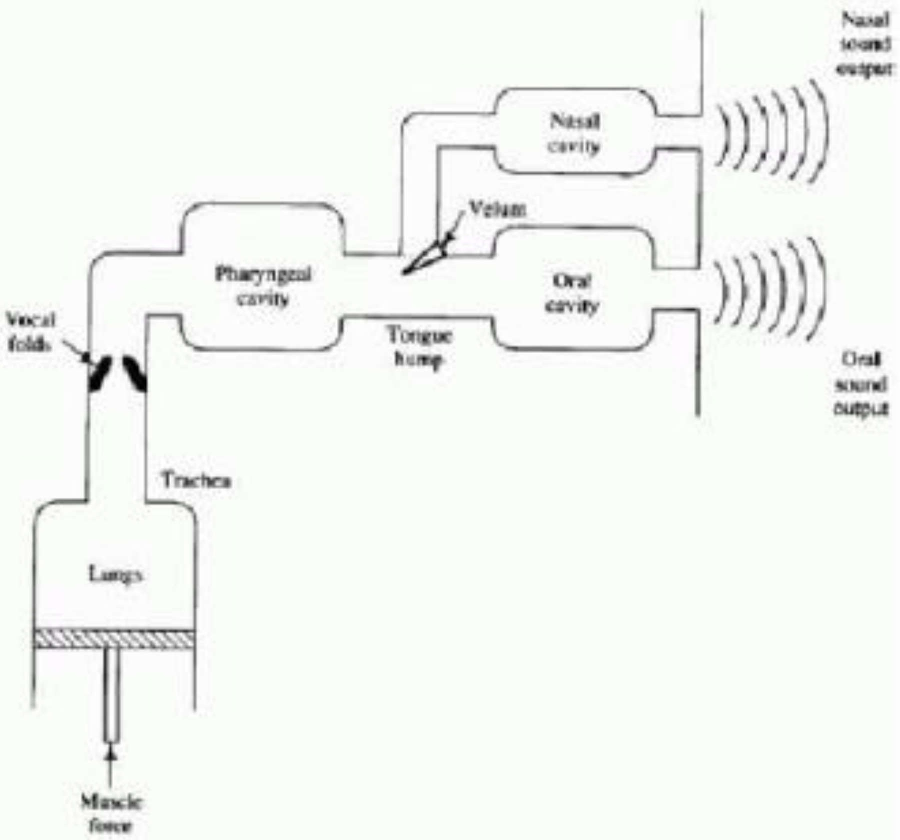
Muhaddisa Barat Ali,
Dept. Of Electrical Engineering, Chalmers University of Technology, Sweden

Mechanism for Human Sound Production



Mechanism for Human Sound Production

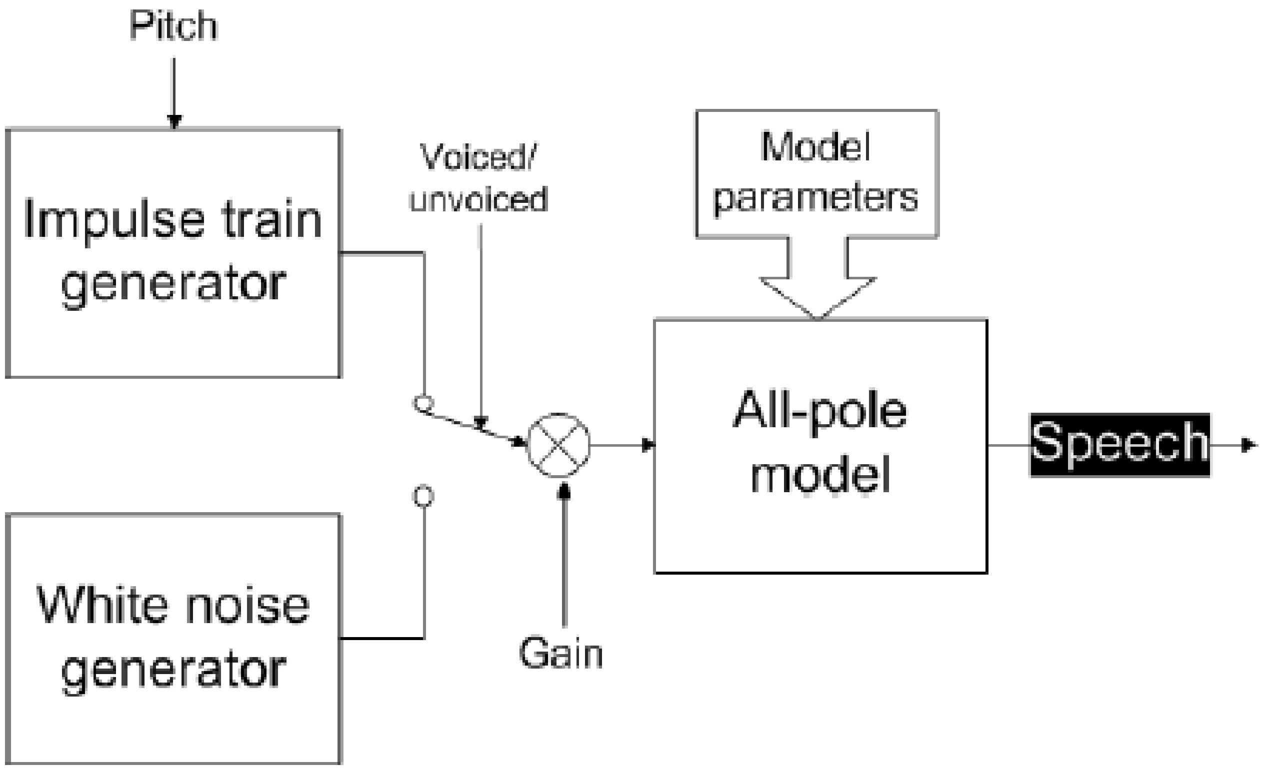




Phonemes

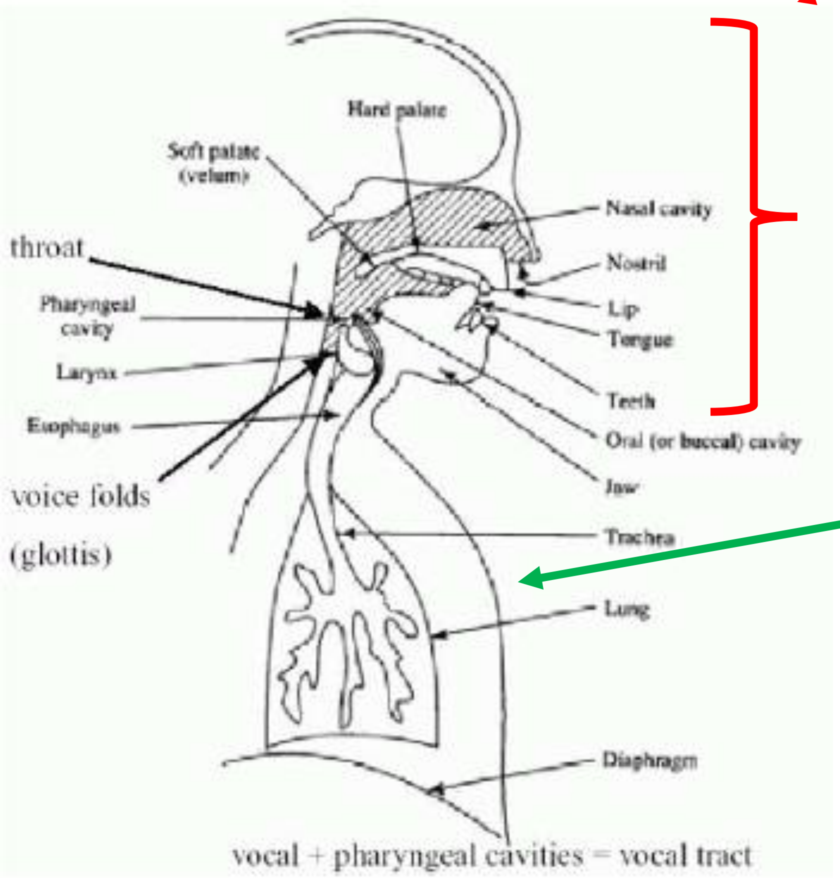
- Phonemes
  - Voiced sound (vowels: a, e, i, o, u)
  - Unvoiced sound (consonants: t, h, p …)

A Mathematical Model for Speech Production

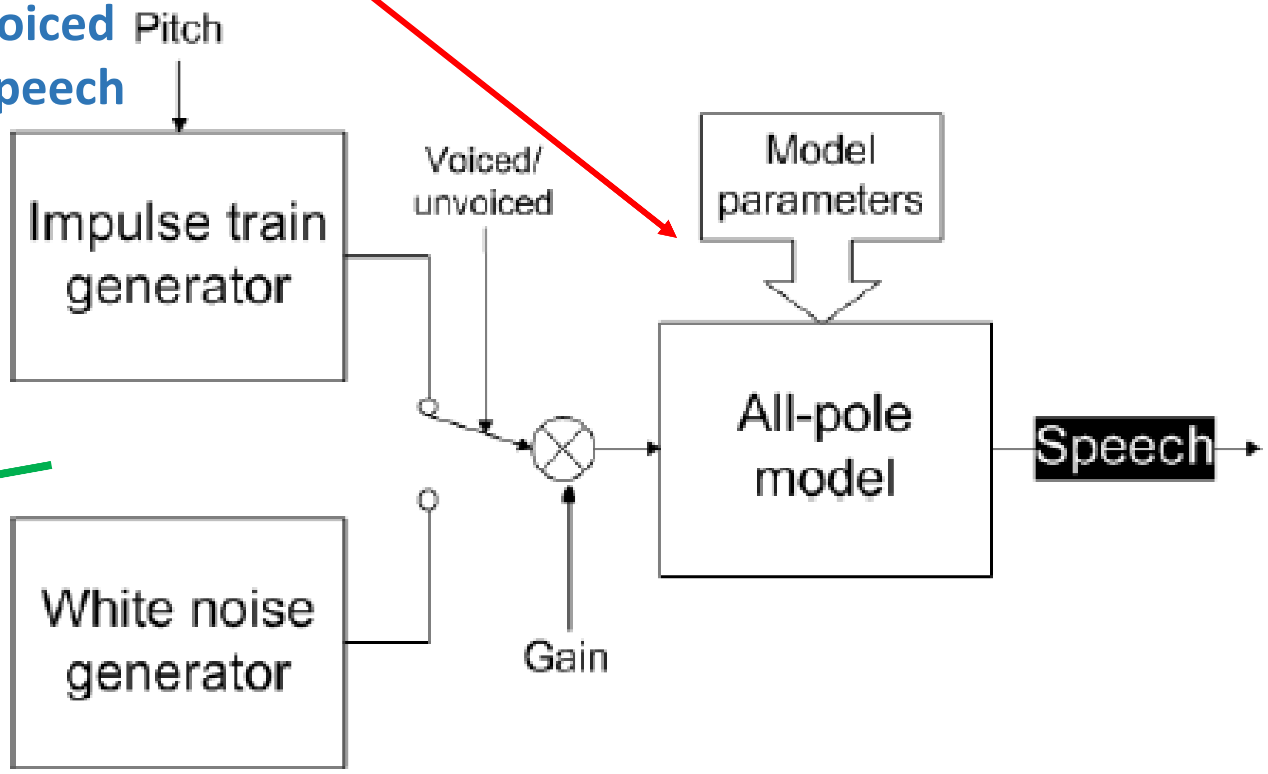


A Mathematical Model for Speech Production

Vocal tract

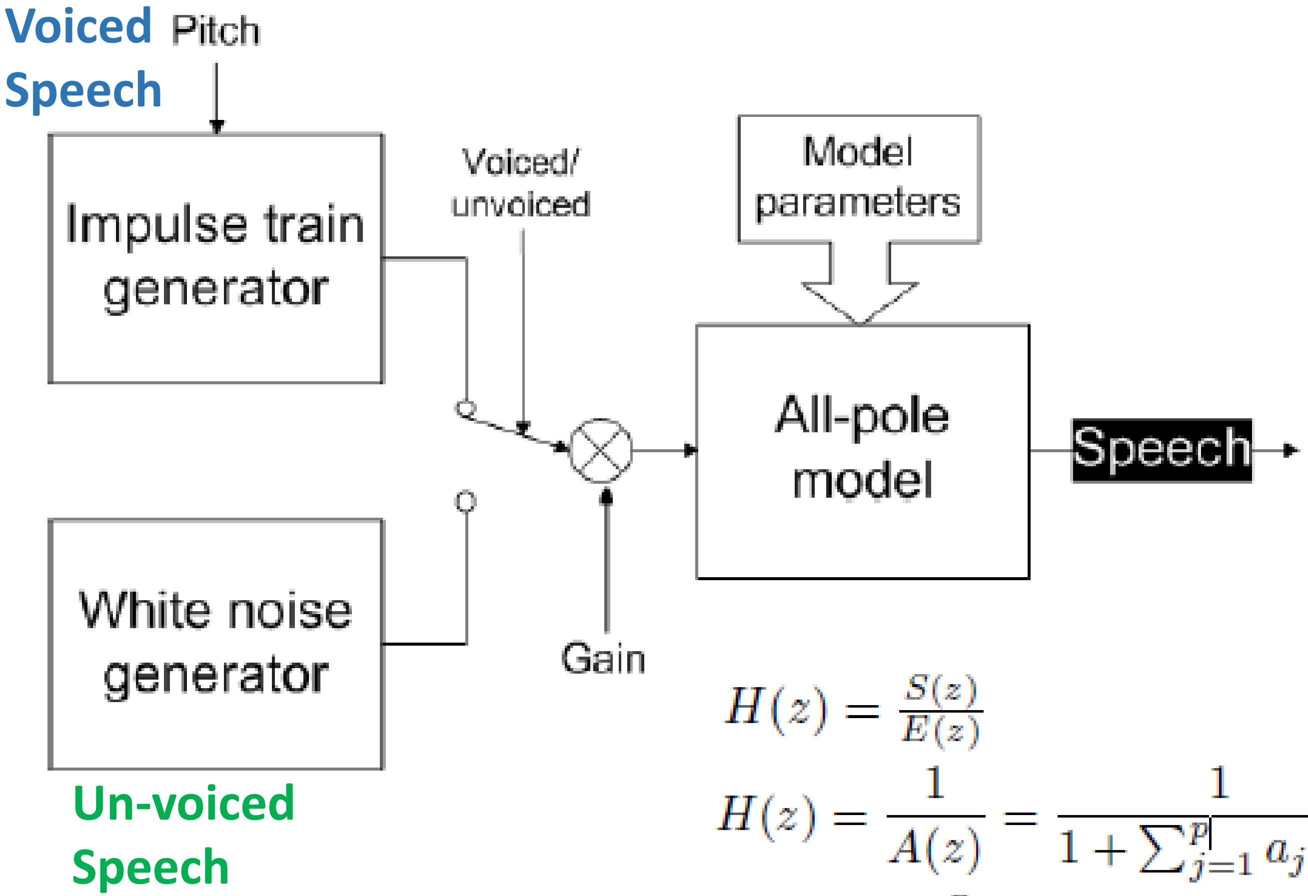


Voiced Speech



Un-voiced Speech

A Mathematical Model for Speech Production



$$H(z) = \frac{S(z)}{E(z)}$$

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{j=1}^{p} a_j z^{-j}}$$

$a_j = model\ parameter$

p = model order

$$s(n) = -\sum_{j=1}^{p} a_j s(n-j) + e(n)$$

$$S(z) = G \cdot H(z)$$

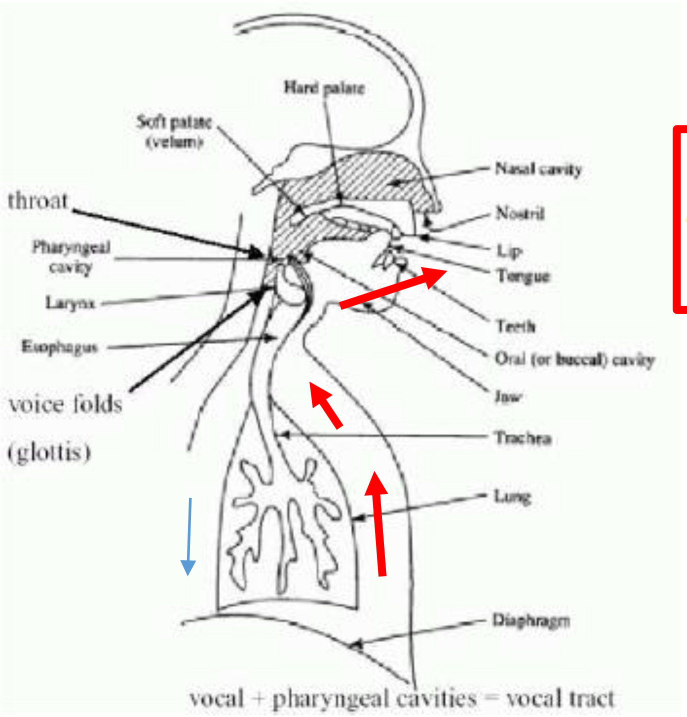
Selecting p

- **Selecting model order p:**

Speech spectrum consists of 3-4 formants (resonant peaks)

Typical choice p = 10

- **Computing the residuals e(n) and the variance σ_e^2(Gain):**

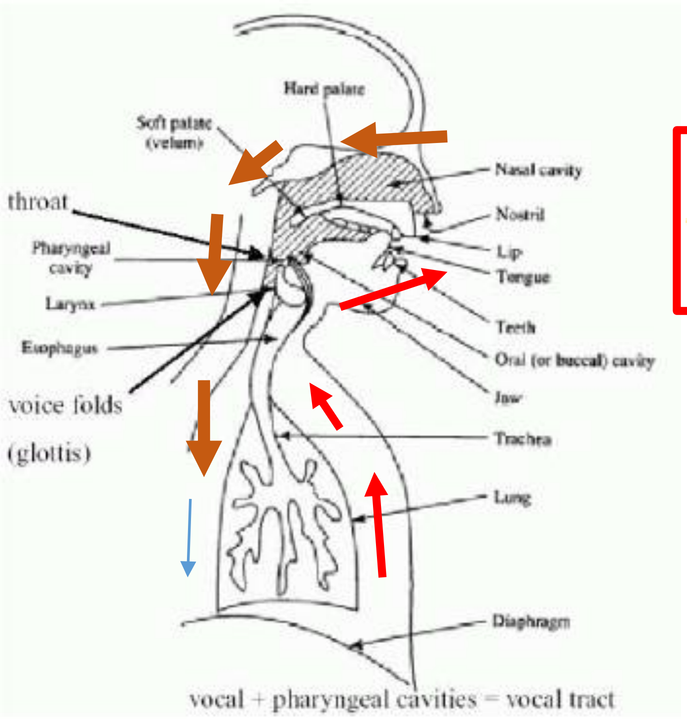


$$\boxed{s(n) = -\sum_{j=1}^{p} a_j s(n-j) + e(n)}$$

- **Computing the residuals e(n) and the variance σ_e^2:**

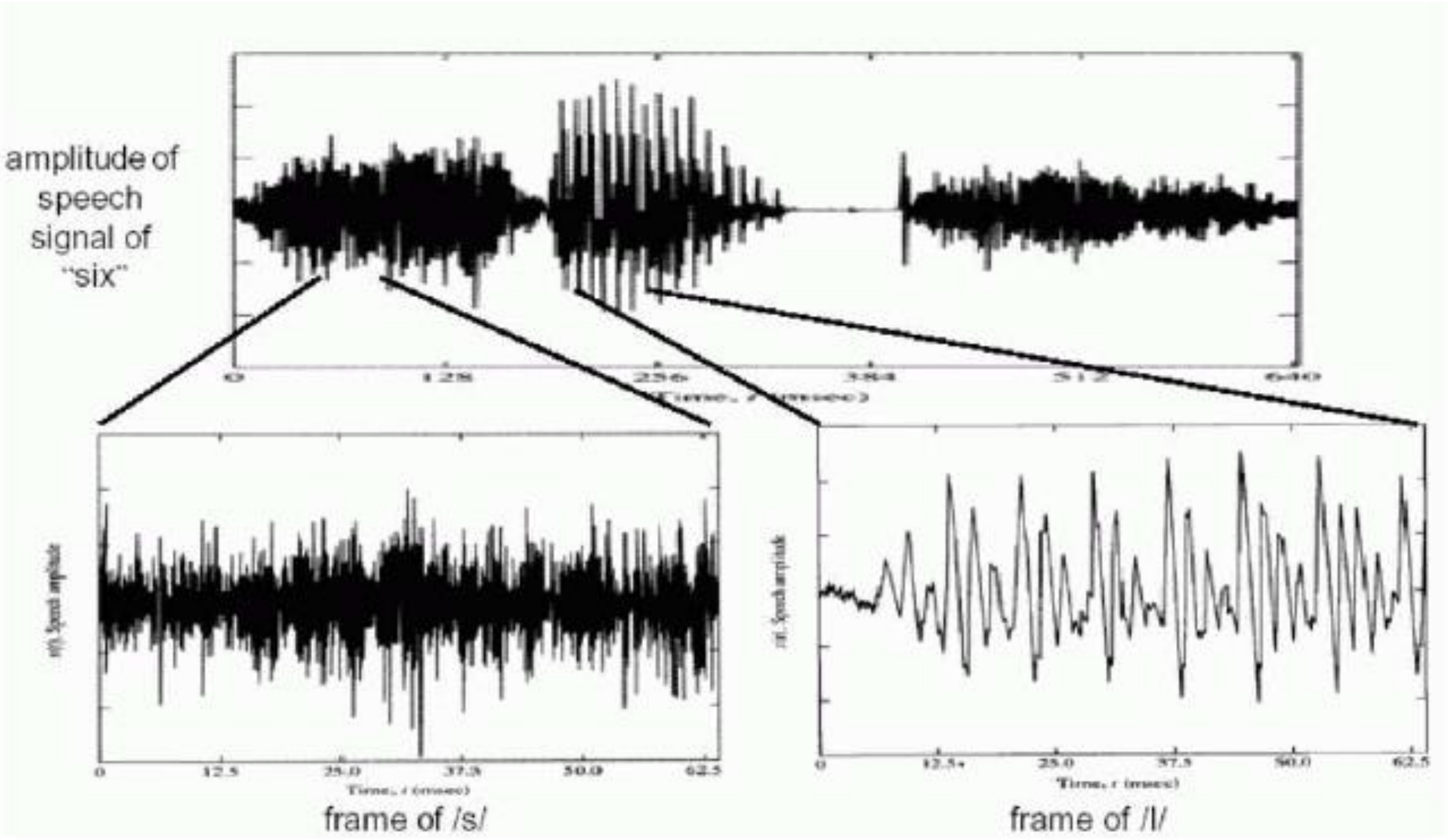
Gain = $|\sigma_e|$

$$\boxed{\hat{e}(n) = s(n) + \sum_{j=1}^{p} \hat{a}_j s(n-j)}$$



$$\boxed{s(n) = -\sum_{j=1}^{p} a_j s(n-j) + e(n)}$$

Vocal cord excitation for voiced and unvoiced



Speech Compression:

- Non-perceivable loss by human ears.

For a 10- 20ms block of speech

- Set of parameters for LPC speech model

$\{a_j \ for \ \ j = 1 \ldots . p = 10, \sigma_e^2$,pitch period$\}$

Speech Compression:

- Non-perceivable loss by human ears.

For a 10-20ms block of speech

- Set of parameters for LPC speech model

$\{a_j \ for \ \ j = 1 \ldots . p = 10, \sigma_e^2$,pitch period$\}$

If each parameter is encoded using 10-bits, total bits to be transmitted will be 12*10=120 bits

- For direct encoding, first sample at 8kHz that gives 160 samples. Convert to bits 160*8= 1280 bits.

Speech Compression:

- Non-perceivable loss by human ears.

For a 10-20ms block of speech

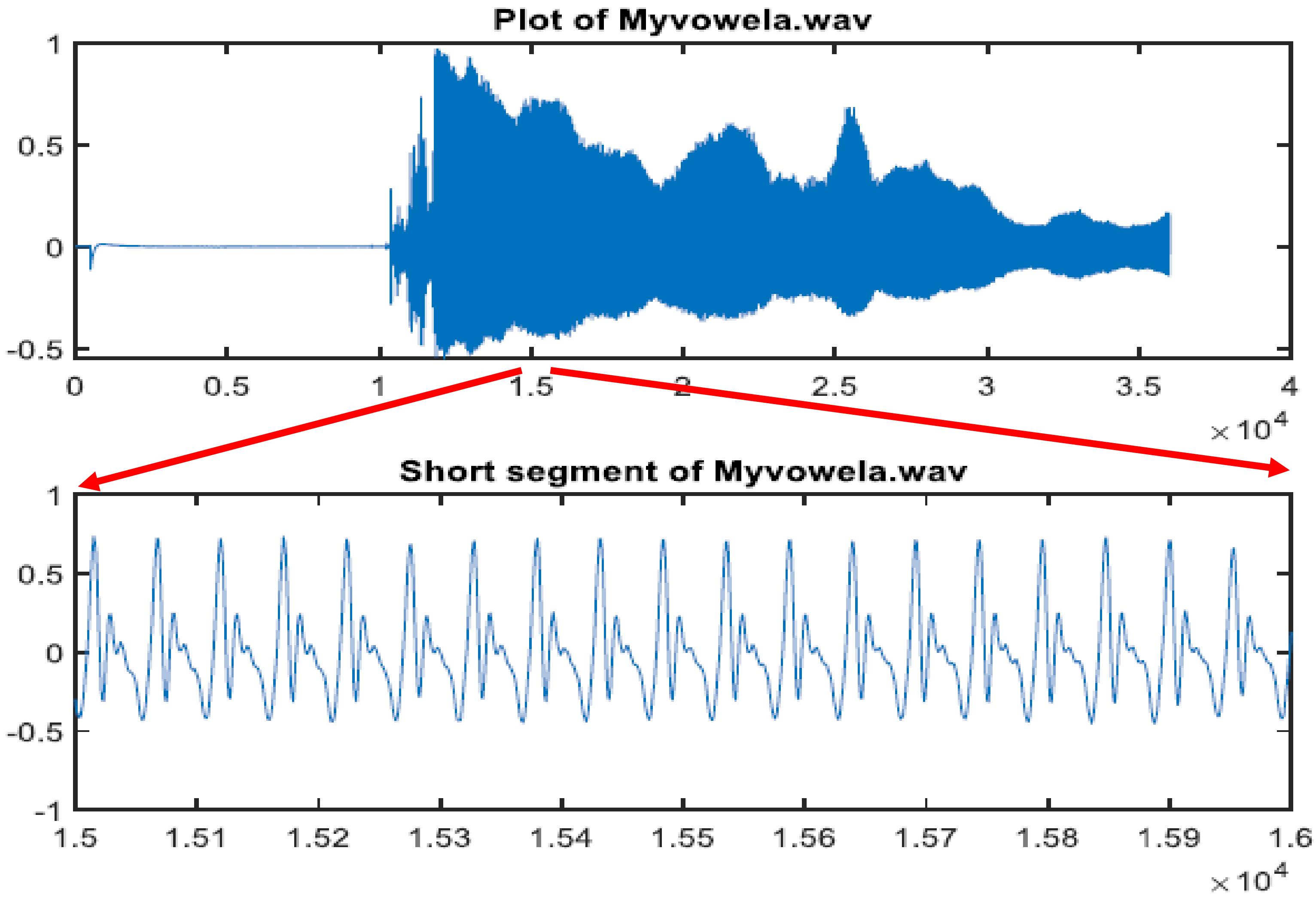
- Set of parameters for LPC speech model

$\{a_j \; for \;\; j = 1 \ldots. p = 10, \sigma_e^2, \text{pitch period}\}$

If each parameter is encoded using 10-bits, total bits to be transmitted will be 12*10=120 bits

- For direct encoding, first sample at 8kHz that gives 160 samples. Convert to bits 160*8= 1280 bits.
- **LPC gives 10 times compression.**

Block based processing of Speech Signal



Model a vocal tract by an all pole-filter for each block.

$$\boxed{\hat{e}(n) = s(n) + \sum_{j=1}^{p} \hat{a}_j^{(i)} s(n-j)}$$

Re-synthesize the Speech

- Use LPC residuals as the input of the filter:

$$\boxed{\hat{s}(n) = -\sum_{j=1}^{p} \hat{a}_j^{(i)} \hat{s}(n-j) + \hat{e}(n)}$$

- Signal $\rightarrow$ Voiced
- Signal $\rightarrow$ Unvoiced

Re-synthesize the Speech

- Use LPC residuals as the input of the filter:

$$\boxed{\hat{s}(n) = -\sum_{j=1}^{p} \hat{a}_j^{(i)} \hat{s}(n-j) + \hat{e}(n)}$$

- Signal
  - → Voiced → Pitch period
  - → Unvoiced → Variance $\boldsymbol{\sigma_e^2}$

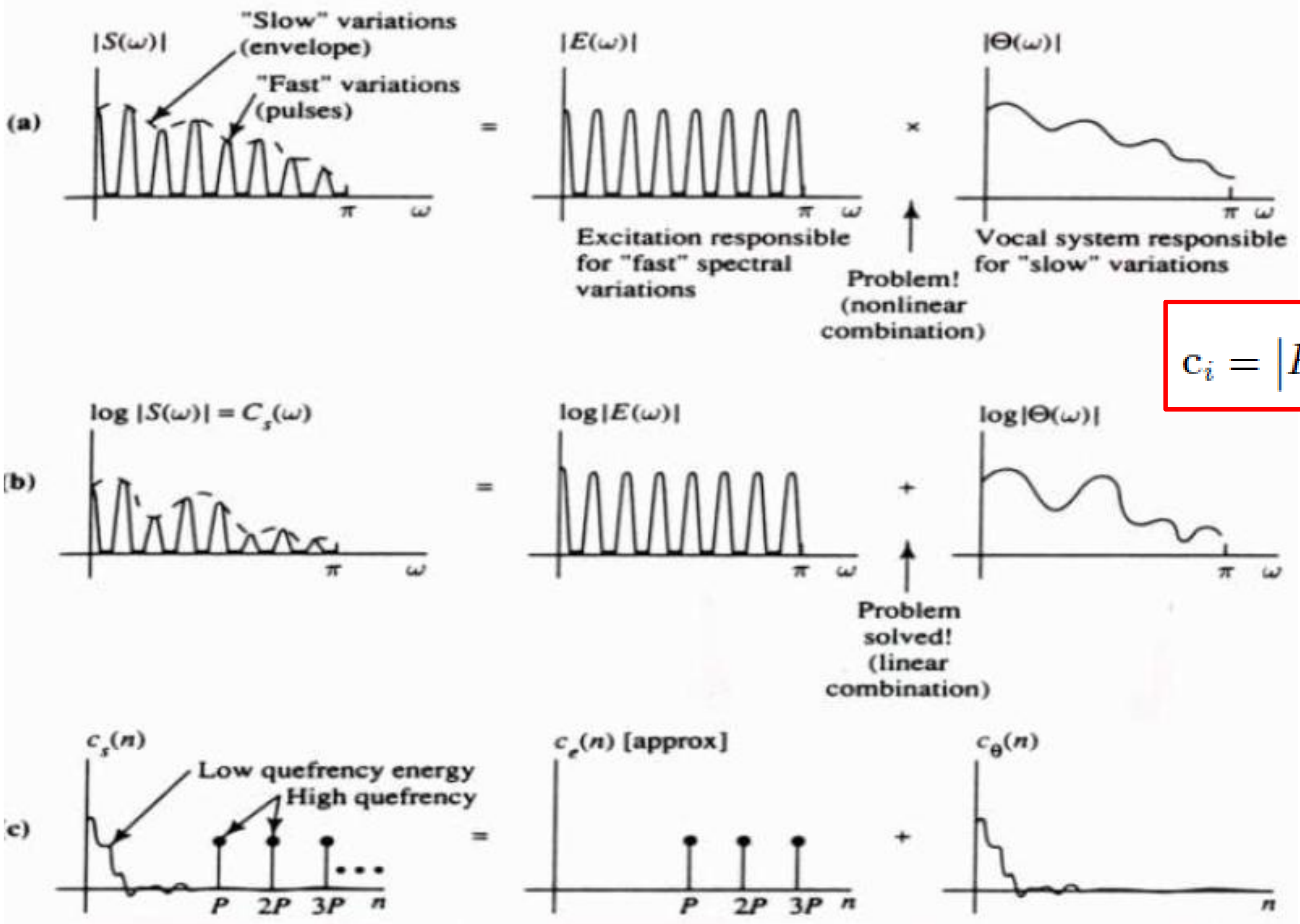
Methods for automatically estimating pitch periods

- Pitch estimation using ACF

$$R_s(\tau) = \sum_{j=(i-1)L+1}^{iL} \hat{e}(j)\hat{e}(j+\tau),$$

- Pitch estimation using Cepstrum domain analysis.

$$\mathbf{c}_i = \left|FFT^{-1}\{log(|FFT(\mathbf{x}_i)|\}\right|$$



$$\mathbf{c}_i = \left|FFT^{-1}\{log(|FFT(\mathbf{x}_i)|\}\right|$$