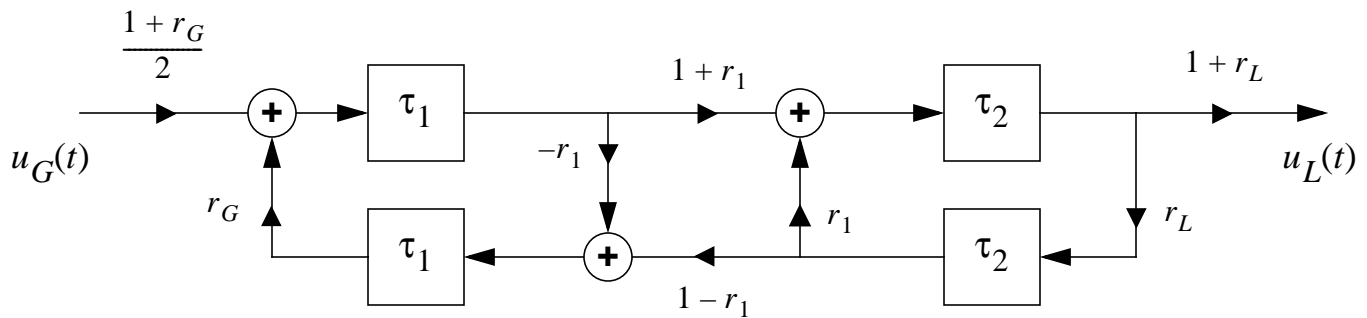
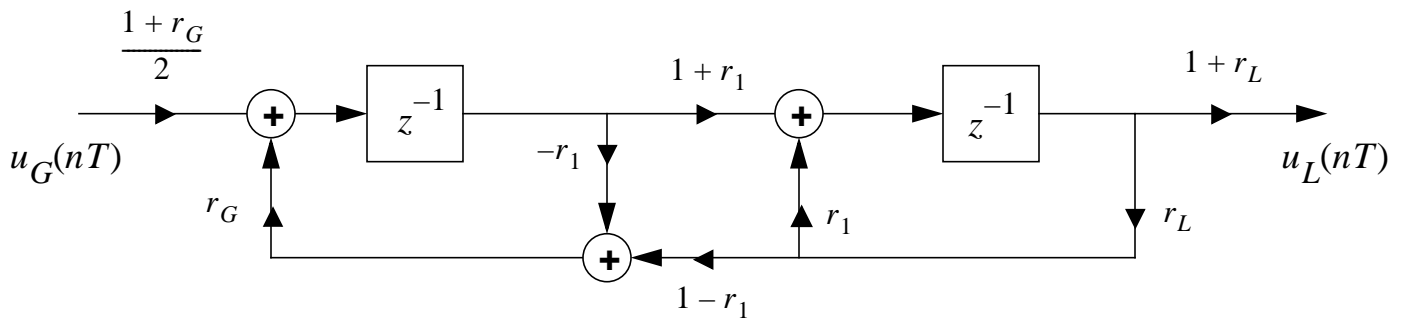


Digital Speech Production Models

Recall our concatenated lossless tube model:



We can approximate this as a digital filter using the sampling theorem:



The transfer function of an N-tube model is:

$$V(z) = \frac{0.5(1+r_G) \prod_{k=1}^N (1+r_k) z^{-N/2}}{D(z)}$$

where

$$D(z) = \begin{bmatrix} 1 & -r_G \end{bmatrix} \begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & -r_N \\ -r_N z^{-N} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

We can compute $D(z)$ recursively:

$$D_o(z) = 1$$

$$D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1}) \quad k = 1, 2, \dots, N$$

$$D(z) = D_N(z)$$



Alternate Digital Filter Implementations Using Digital Resonators

Note that for $D(z)$ to have real coefficients, zeros must occur in complex conjugate pairs. We can transform zeros in the Laplace domain:

$$s_k, s_k^* = -\sigma \pm j2\pi F_k$$

The corresponding complex conjugate poles in the discrete-domain are:

$$\begin{aligned} z_k, z_k^* &= e^{-\sigma_k T} e^{\pm j2\pi F_k T} \\ &= e^{-\sigma_k T} \cos(2\pi F_k T) \pm j e^{-\sigma_k T} \sin(2\pi F_k T) \end{aligned}$$

Note that magnitude of the pole in the z -plane is related to the bandwidth.

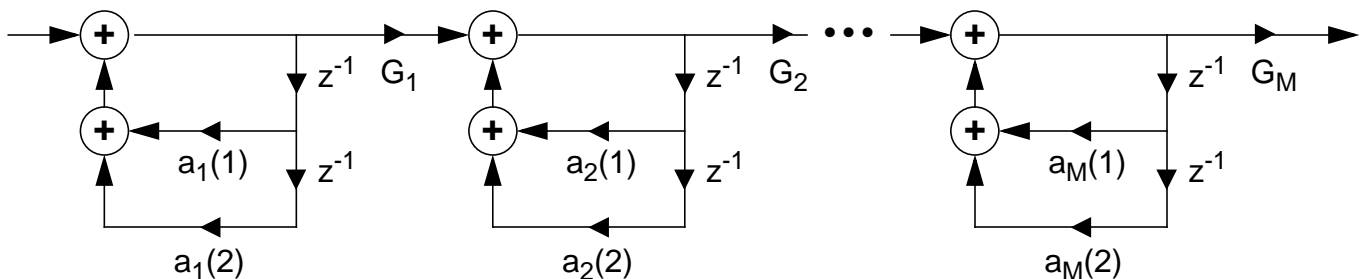
We can write a transfer function as a product of these poles:

$$V(z) = \prod_{k=1}^M V_k(z)$$

where

$$V_k(z) = \frac{(1 - 2|z_k| \cos(2\pi F_k T) + |z_k|^2)}{(1 - 2|z_k| \cos(2\pi F_k T)z^{-1} + |z_k|^2 z^{-2})}$$

This is an all-pole filter. It can be realized using a number of structures:
Under what conditions is this filter stable?



where,

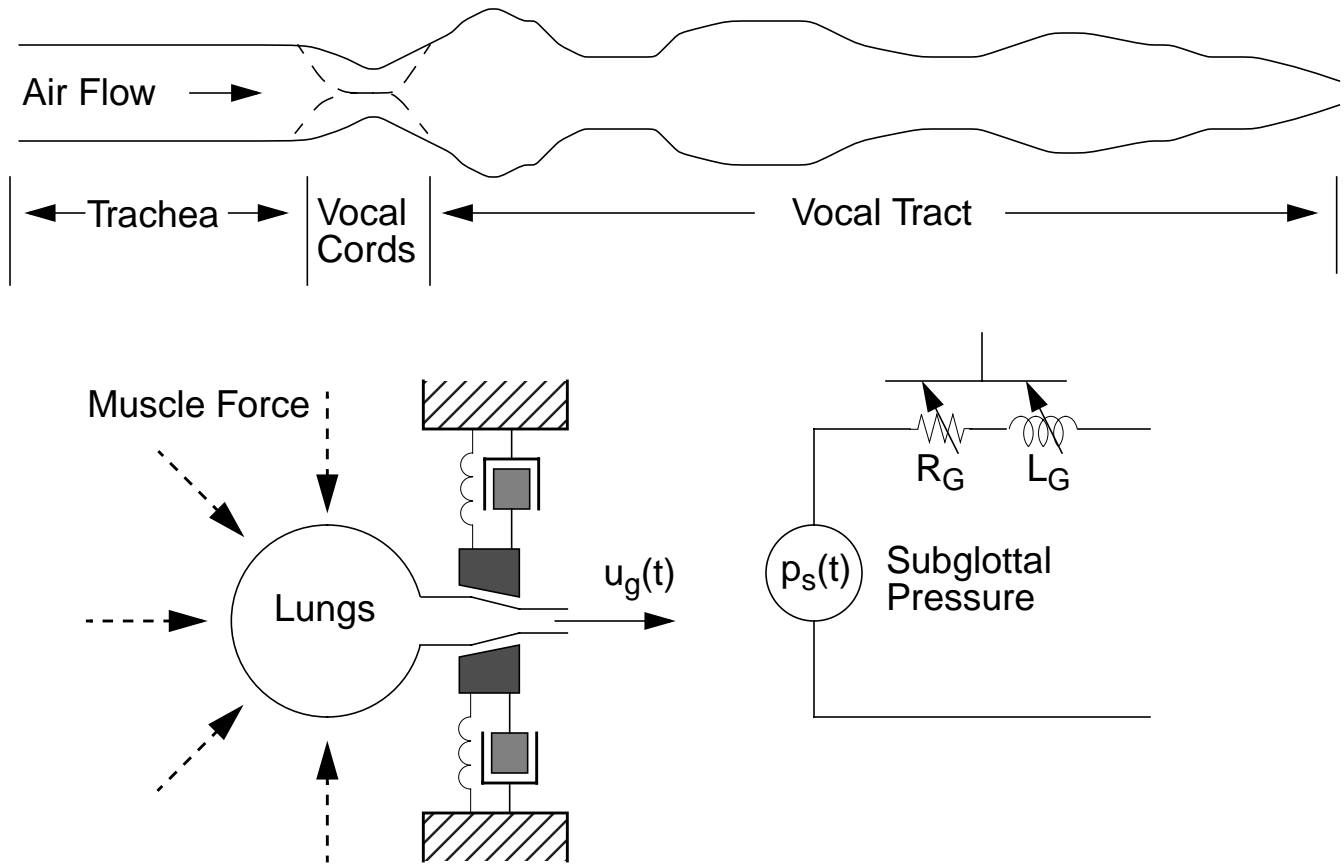
$$V_k(z) = \frac{G_k}{1 - a_k(1)z^{-1} - a_k(2)z^{-2}}$$

$$a_k(1) = 2|z_k| \cos(2\pi F_k T) \quad a_k(2) = -|z_k|^2 \quad G_k = 1 - 2|z_k| \cos(2\pi F_k T) + |z_k|^2$$



Excitation Models

How do we couple energy into the vocal tract?



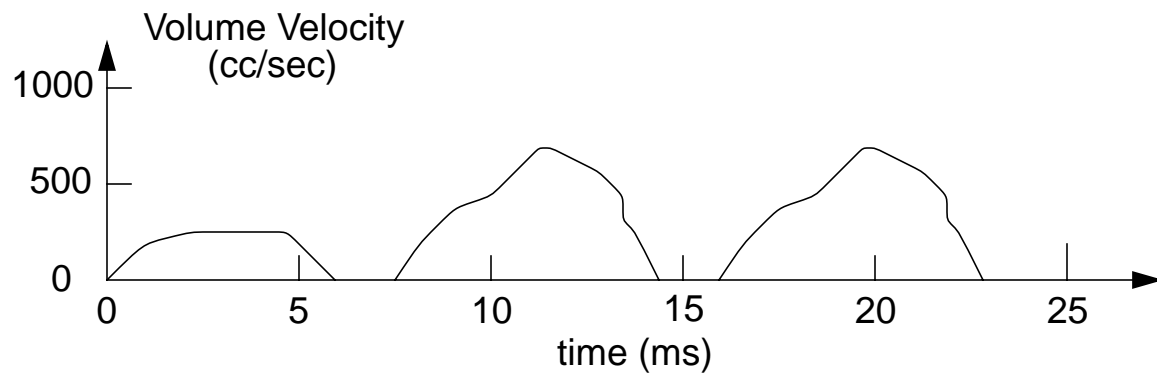
The glottal impedance can be approximated by:

$$Z_G = R_G + j\Omega L_G$$

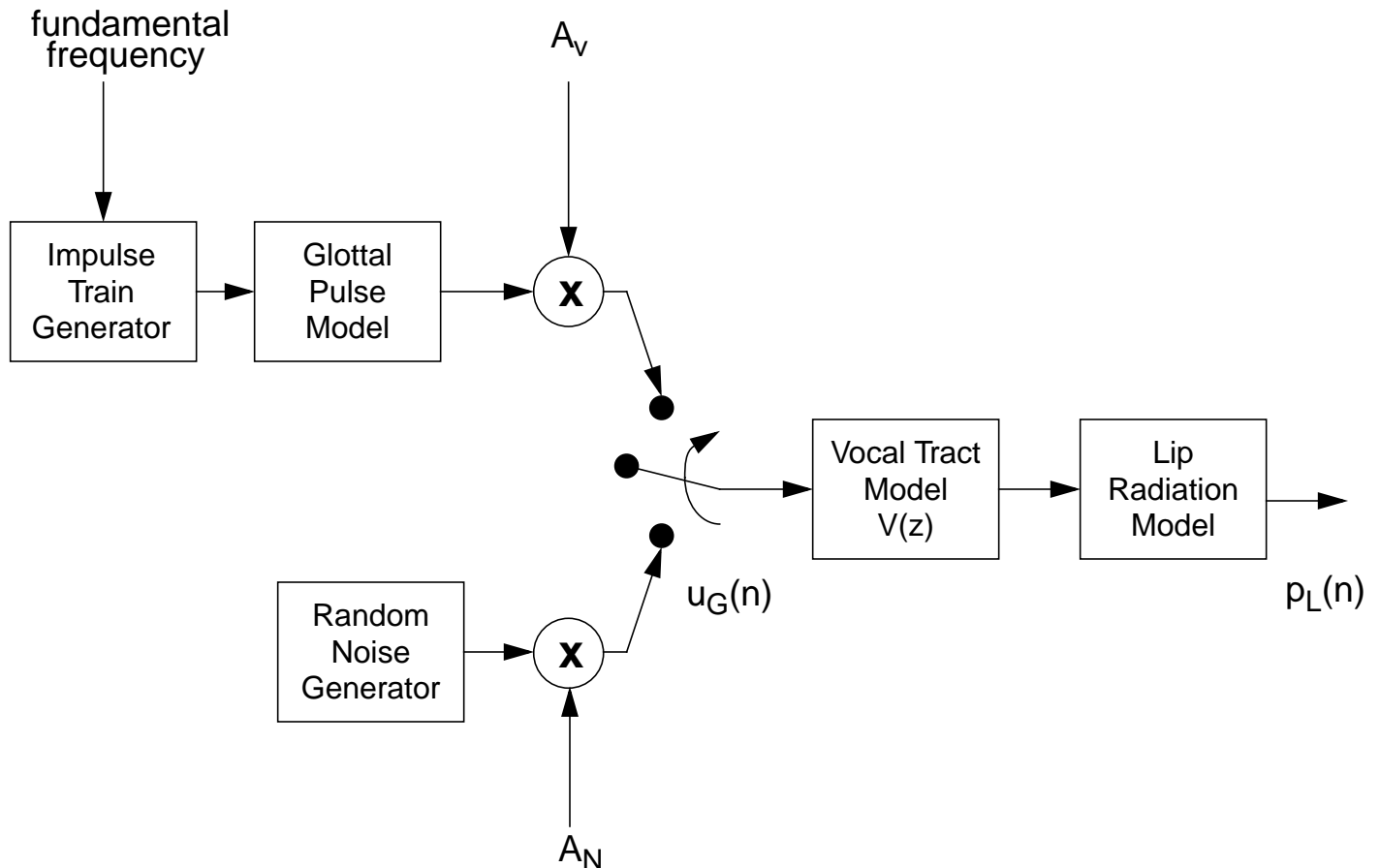
The boundary condition for the volume velocity is:

$$U(0, \Omega) = U_G(\Omega) - P(0, \Omega) / Z_G(\Omega)$$

For voiced sounds, the glottal volume velocity looks something like this:



The Complete Digital Model (Vocoder)



Notes:

- Sample frequency is typically 8 kHz to 16 kHz
- Frame duration is typically 10 msec to 20 msec
- Window duration is typically 30 msec
- Fundamental frequency ranges from 50 Hz to 500 Hz
- Three resonant frequencies are usually found within 4 kHz bandwidth
- Some sounds, such as sibilants ("s") have extremely high bandwidths

Questions:

What does the overall spectrum look like?

What happened to the nasal cavity?

What is the form of $V(z)$?

Linear Prediction

How do we estimate the vocal tract parameters?

Recall our digital filter model:

$$A(z) = 1 - \sum_{k=1}^p a_k z^{-k}$$

This corresponds to a finite difference equation of the form:

$$y(n) = a_1 y(n-1) + a_2 y(n-1) + \dots + a_p y(n-p) + x(n)$$

We predict the current value, $y(n)$, based on its previous values and the new input value — this is known as linear prediction.

We can define the energy of the prediction error as:

$$E(n) = \sum_{n=0}^{N-1} [y(n) - \tilde{y}(n)]^2$$

where $\tilde{y}(n)$ is the predicted value. We can derive an equation for the computation of $\{a_k\}$ by minimizing the mean-square error (differentiate the energy of the error w.r.t. a_i and solve for a_i). This yields:

$$\mathbf{a} = \mathbf{R}^{-1} \mathbf{r}$$

where:

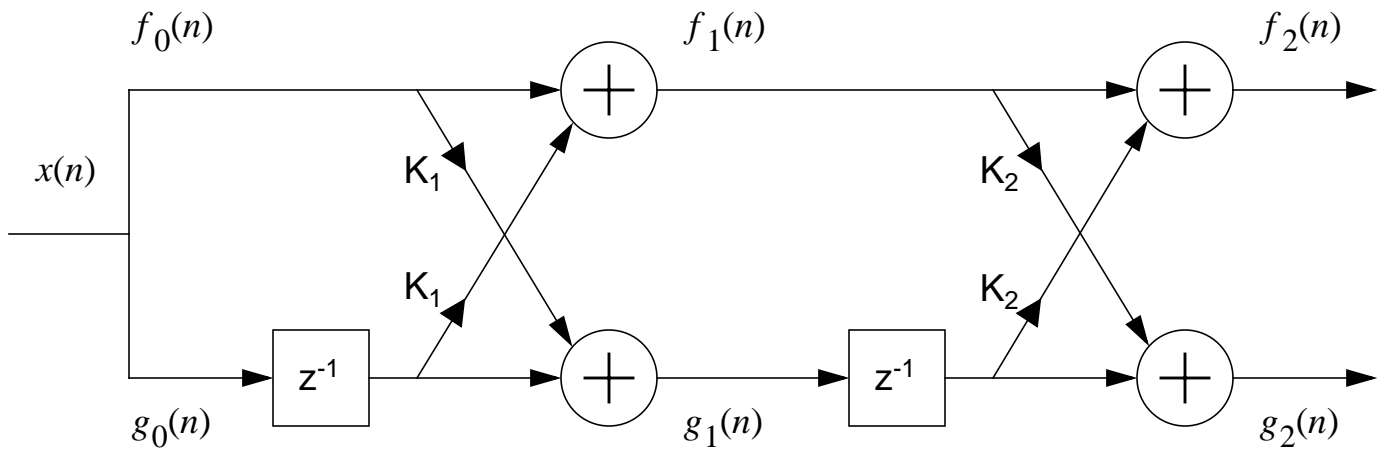
$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} R(0) & R(1) & \dots & R(p-1) \\ R(1) & R(0) & \dots & R(p-2) \\ \dots & \dots & \dots & \dots \\ R(p-1) & R(p-2) & \dots & R(0) \end{bmatrix} \quad \mathbf{r} = \begin{bmatrix} R(1) \\ R(2) \\ \dots \\ R(p) \end{bmatrix}$$

and,

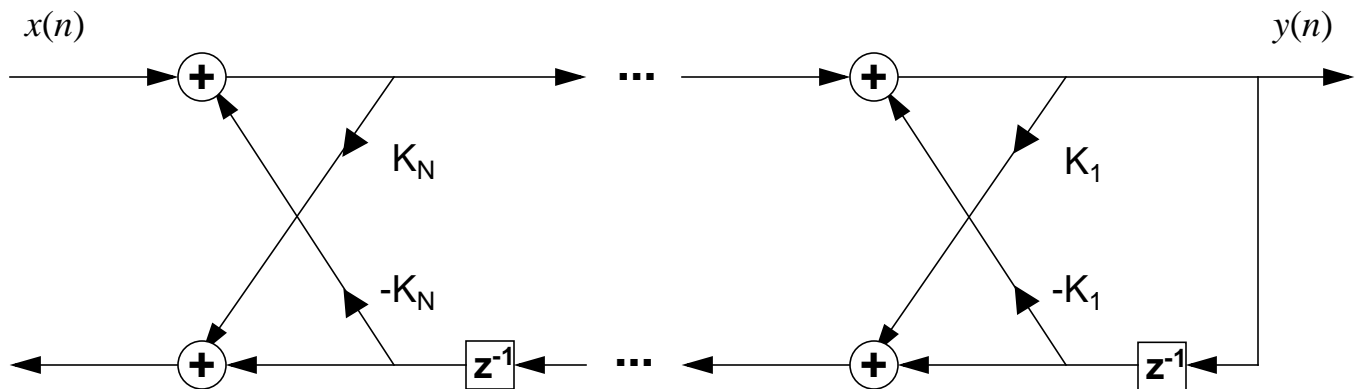
$$R(i) = \sum_{n=0}^{N-i} x(n)x(n-i).$$

Relationship to the Lattice Filters and Reflection Coefficients

The standard direct-form FIR filter can be implemented in a lattice structure:



The inverse, or Infinite Impulse Response (IIR) equivalent, is an all-pole filter:



The coefficients $\{K_i\}$ are called reflection coefficients, and can be computed directly from the signal:

$$K_i = \frac{\sum_{m=0}^{N-1} f_{i-1}(m)g_{i-1}(m-1)}{\left\{ \left(\sum_{m=0}^{N-1} (f_{i-1}(m))^2 \right) \left(\sum_{m=0}^{N-1} (f_{i-1}(m-1))^2 \right) \right\}^{1/2}}$$

For the filter to be stable, these reflection coefficients must be bounded: $|K_i| \leq 1$.



Transformations Between Parameters

The predictor coefficients, reflection coefficients, and area ratios represent alternate descriptions of the same information:

Predictor to reflection coefficient transformation:

for $i = p, p-1, \dots, 1$

$$k_i = a_i(i)$$

$$a_{i-1}(j) = \frac{a_i(j) + k_i a_i(i-j)}{1 - k_i^2} \quad 1 \leq j \leq i-1$$

Reflection to predictor coefficient transformation:

for $i = 1, 2, \dots, p$

$$a_i(i) = k_i$$

$$a_i(j) = a_{i-1}(j) - k_i a_{i-1}(i-j) \quad 1 \leq j \leq i-1$$

Durbin Recursion: an efficient algorithm to solve linear equations involving symmetric matrices):

for $i = 1, 2, \dots, p$

$$E_0 = R(0)$$

$$k_i = \left(R(i) - \sum_{j=1}^{i-1} a_{i-1}(j) R(i-j) \right) / E_{i-1}$$

for $j = 1, 2, \dots, i-1$

$$a_i(i) = k_i$$

$$a_i(j) = a_{i-1}(j) - k_i a_{i-1}(i-j)$$

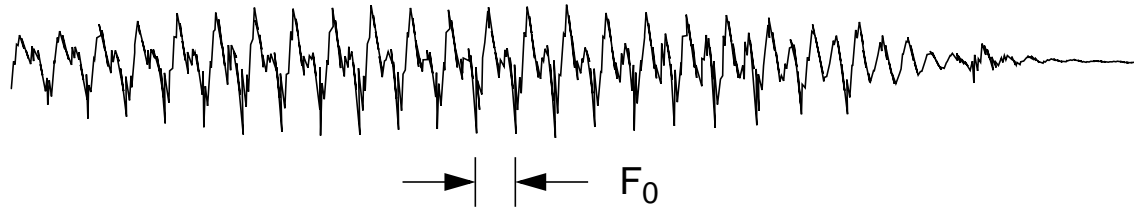
$$E_i = (1 - k_i^2) E_{i-1}$$

Log of the ratio of the areas of adjacent sections of a lossless tube:

$$g_i = \log \left[\frac{A_{i+1}}{A_i} \right] = \log \left[\frac{1 - k_i}{1 + k_i} \right] \quad 1 \leq i \leq p$$

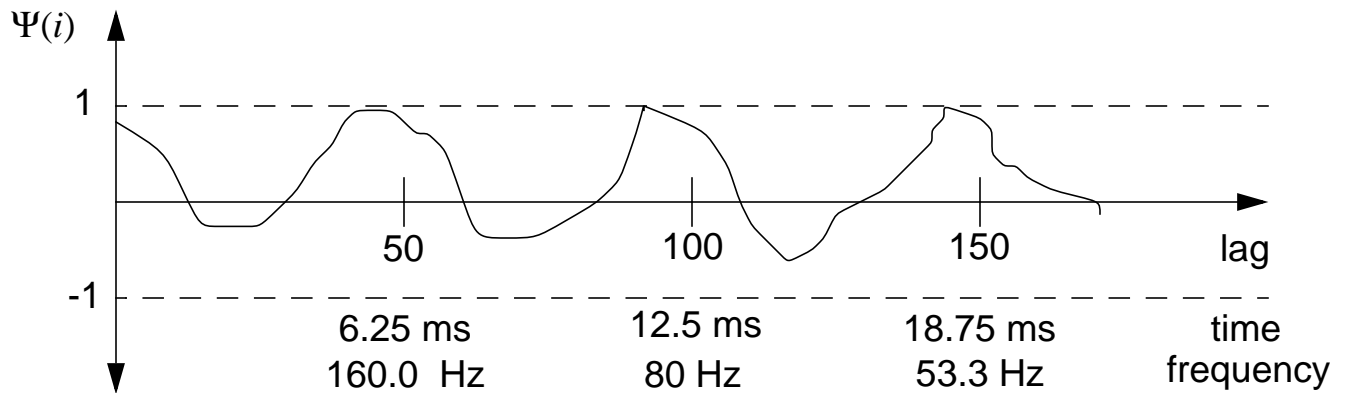
Fundamental Frequency Analysis

How do we determine the fundamental frequency?



We use the (statistical) autocorrelation function:

$$\Psi(i) = \frac{\sum_{n=0}^{N-1} x(n)x(n-i)}{\sqrt{\left(\sum_{n=0}^{N-1} x(n)^2\right)\left(\sum_{n=0}^{N-1} x(n-i)^2\right)}}$$



Other common representations:

Average Magnitude Difference Function (AMDF):

$$\gamma(i) = \sum_{n=0}^{N-1} |x(n) - x(n-i)|$$

Zero Crossing Rate:

$$F_0 = \frac{ZF_s}{2} \quad Z = \sum_{n=0}^{N-1} |\text{sgn}[x(n)] - \text{sgn}[x(n-1)]|$$