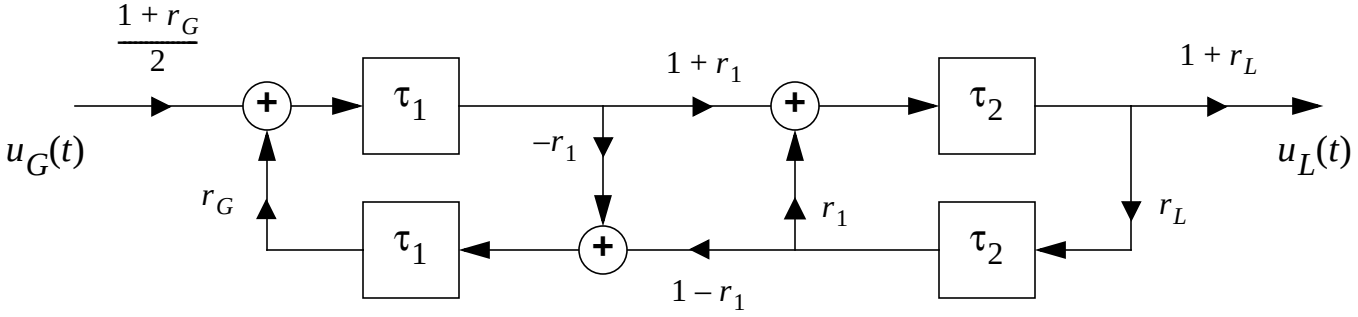
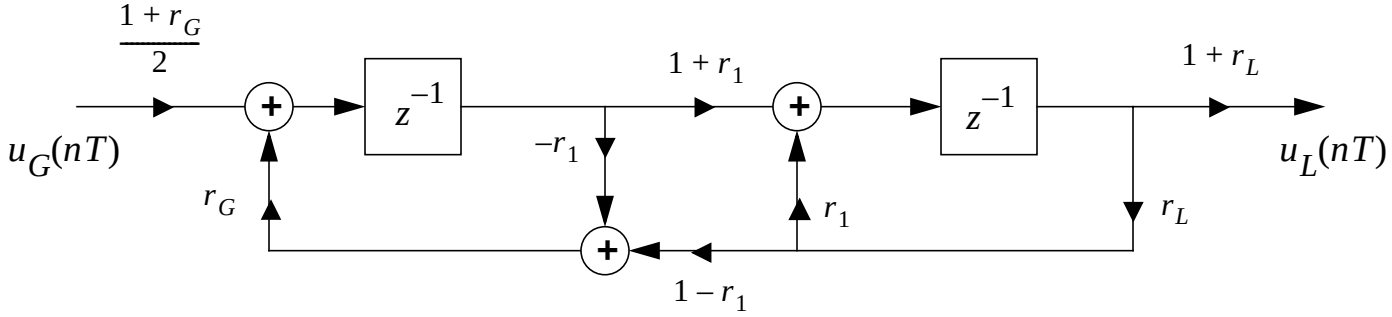
Digital Speech Production Models

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:

$$D_o(z) = 1$$

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

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

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:

$$z_k, z_k^* = e^{-\sigma_k T} e^{\pm j2\pi F_k T}$$

$$= e^{-\sigma_k T}\cos(2\pi F_k T) \pm j e^{-\sigma_k T}\sin(2\pi F_k T)$$

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

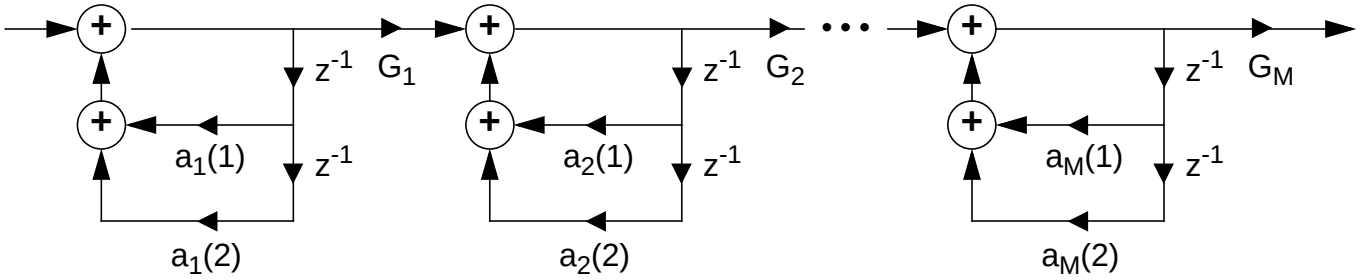
We can write a transfer function as a product of these poles:

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

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?



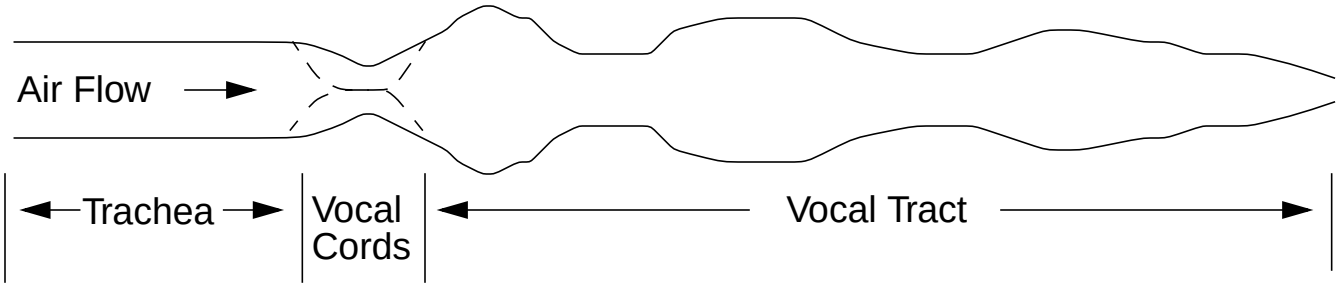
where,

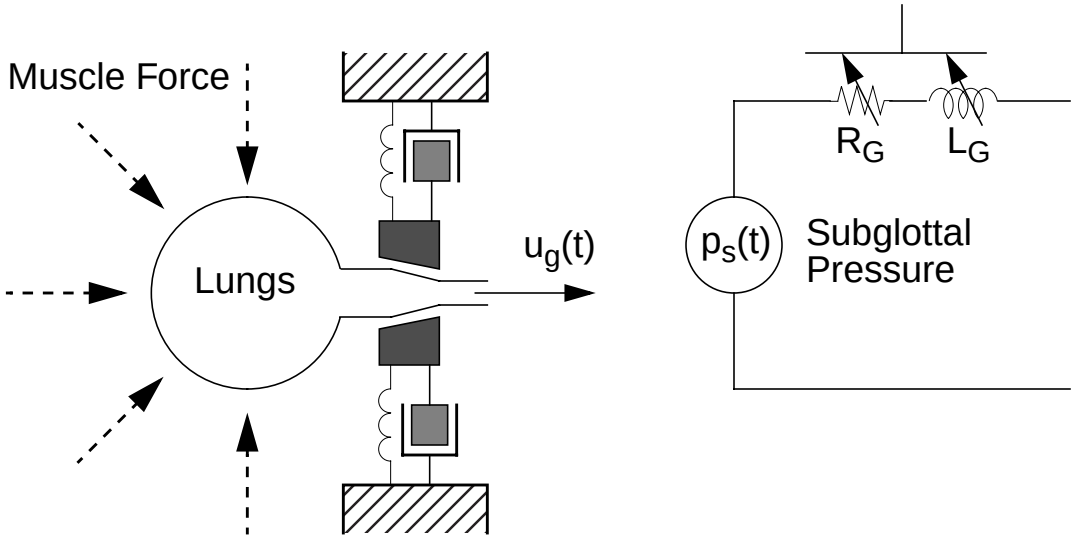
$$V_k(z) = \frac{G_M}{1 - a_k(1)z^{-1} - a_k(2)z^{-2}}$$

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

Excitation Models

How do we couple energy into the vocal tract?





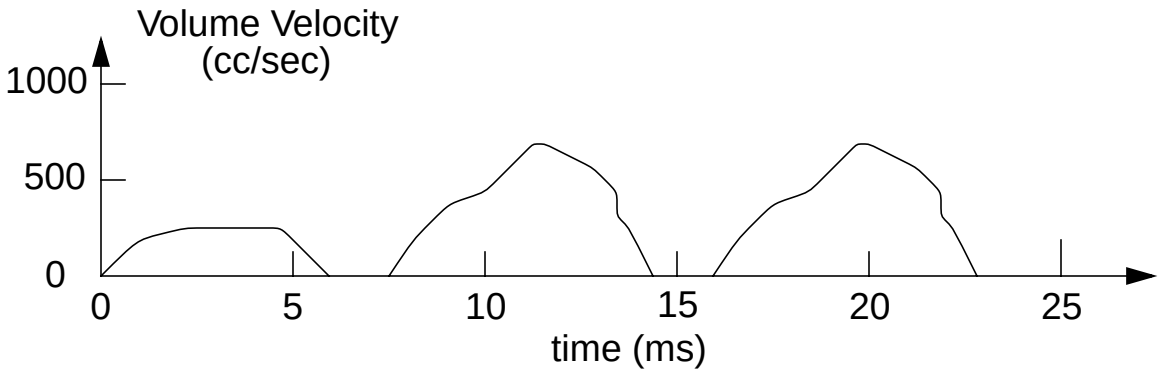
The glottal impedance can be approximated by:

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

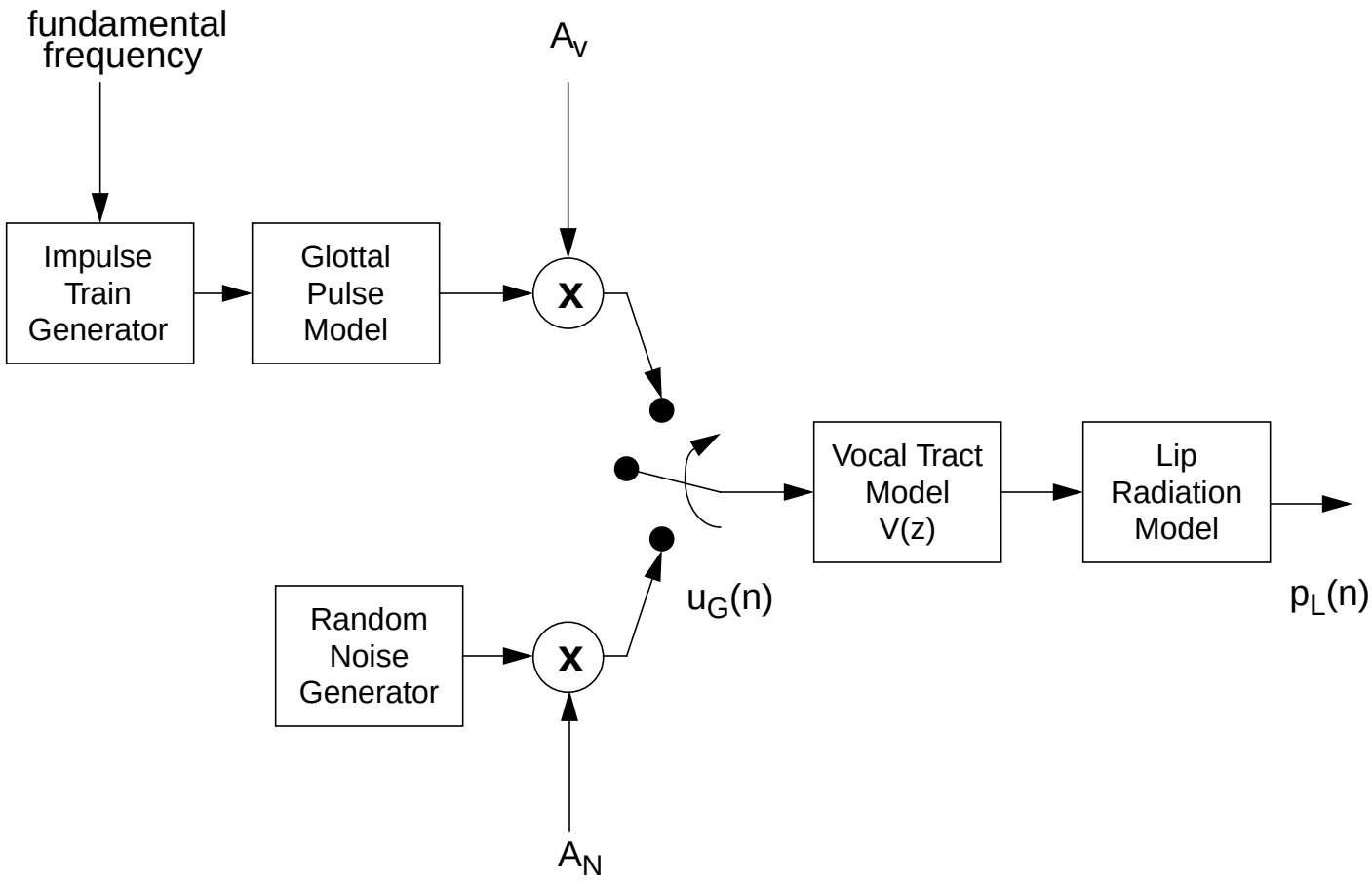
The boundary condition for the volume velocity is:

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

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:

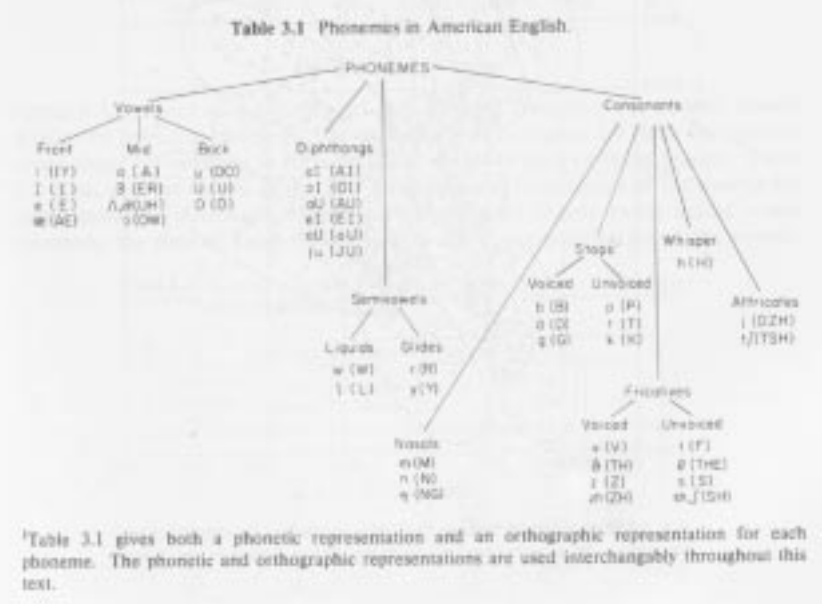
What does the overall spectrum look like?

What happened to the nasal cavity?

What is the form of V(z)?

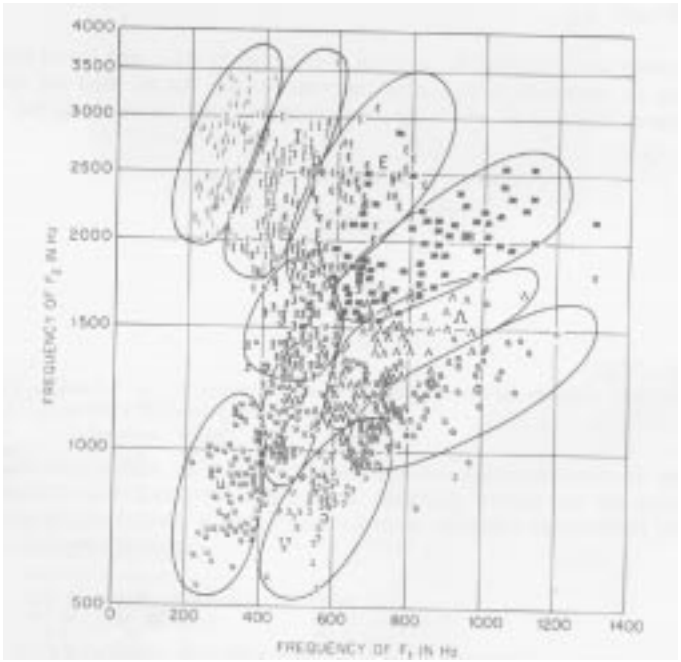
The Vowel Triangle

Each fundamental speech sound can be categorized according to the position of the articulators. This is known as the study of Acoustic Phonetics.



[1]Table 3.1 gives both a phonetic representation and an orthographic representation for each phoneme. The phonetic and orthographic representations are used interchangably throughout this text.

We can characterize a vowel sound by the locations of the first and second spectral resonances, known as a formant frequencies:



Some voiced sounds, such as diphthongs, are transitional sounds that move from one vowel location to another.

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:

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

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:

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

where:

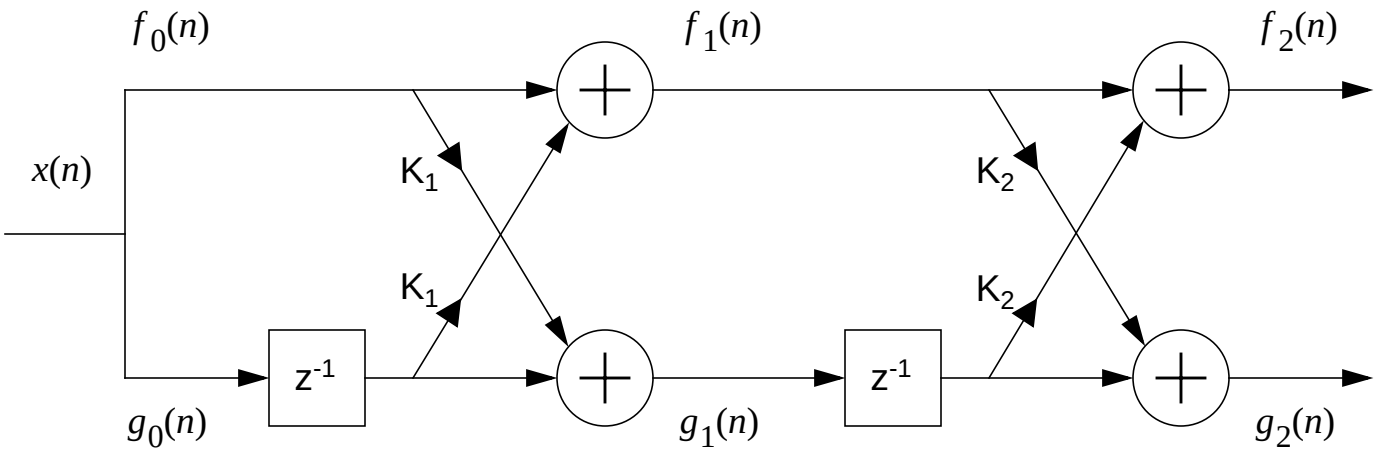
$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_p \end{bmatrix} \qquad \boldsymbol{R} = \begin{bmatrix} R(0) & R(1) & \ldots & R(p-1) \\ R(1) & R(0) & \ldots & R(p-2) \\ \ldots & \ldots & \ldots & \ldots \\ R(p-1) & R(p-2) & \ldots & R(0) \end{bmatrix} \qquad \boldsymbol{r} = \begin{bmatrix} R(1) \\ R(2) \\ \ldots \\ R(p) \end{bmatrix}$$

and,

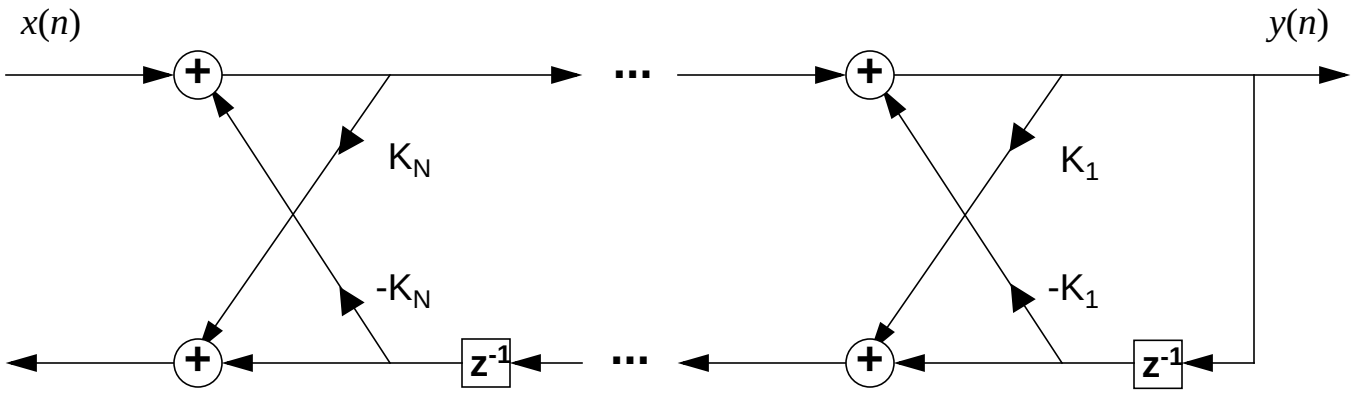
$$R(i) = \sum_{n=0}^{N-i} x(n)x(n-i).$$

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:

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

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:

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

$$k_i = a_i(i)$$

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

Reflection to predictor coefficient transformation:

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

$$a_i(i) = k_i$$

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

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

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

$$E_0 = R(0)$$

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

$$for\ \mathrm{j} = 1, 2, \ldots, i-1$$

$$a_i(i) = k_i$$

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

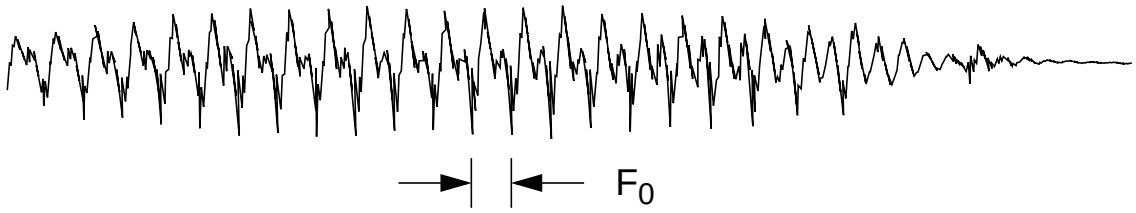
$$E_i = (1 - k_i^2) E_{i-1}$$

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

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

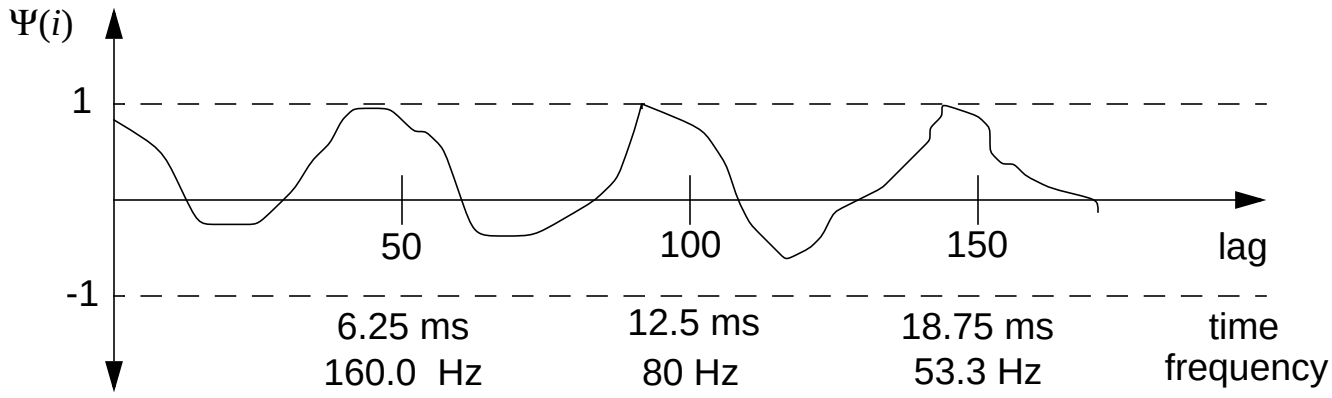
Fundamental Frequency Analysis

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:

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

How Do We Compare Two Models?

(1) **Log-Likelihood:**

$$D(\boldsymbol{a}, \hat{\boldsymbol{a}}) = \frac{\boldsymbol{a} R_{\hat{\boldsymbol{a}}} \boldsymbol{a}^{\dagger}}{\hat{\boldsymbol{a}} R_{\hat{\boldsymbol{a}}} \hat{\boldsymbol{a}}^{\dagger}} = \frac{\left[b(0)\hat{R}(0) + 2 \sum_{i=1}^{p} b(i)\hat{R}(i) \right]}{\hat{\boldsymbol{a}} R_{\hat{\boldsymbol{a}}} \hat{\boldsymbol{a}}^{\dagger}}$$

where

$$b(i) = \sum_{j=0}^{p-i} a(j)a(j+i) \qquad 0 \le i \le p$$

This formulation is no longer used but was of great interest initially because it described the matching problem in a probability space, and it resulted in a computationally-efficient procedure for recognition.

(2) **Euclidean Distance Via Cepstral Coefficients:**

Cepstral coefficients can be computed from predictor coefficients:

$$\hat{h}(n) = a(n) + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) \hat{h}(k)a(n-k) \qquad n \ge 1$$

A meaningful distance metric can be computed as:

$$D(\boldsymbol{h_1}, \hat{\boldsymbol{h_2}}) = \sum_{i=1}^{M} \left(h_1(n) - h_2(n)\right)^2$$

(3) **Statistically Decorrelated Parameters (Principal Components):**

A statistically meaningful distance can be computed by orthogonalizing the parameter space:

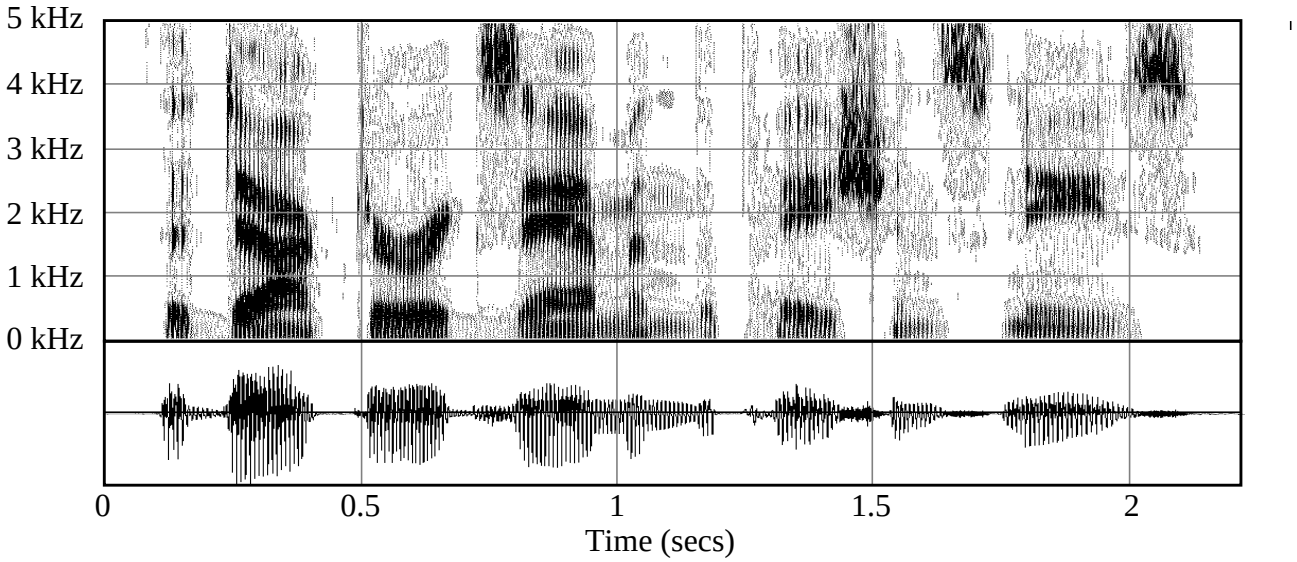
$$b = \Psi(a - \mu)$$

where Ψ is a whitening transformation:

$$\Psi = \Lambda^{-1/2} \Phi^{\dagger}$$

and Λ denotes a diagonal matrix of eigenvalues, and Φ denotes a matrix of eigenvectors. The net result of this transformation is to produce a space in which all dimensions have equal importance.

The Spectrogram



Why do we use the spectrogram to visualize speech data?

The spectrogram is computed by a sliding window approach based on a Discrete Fourier Transform (DFT):

$$X(f) = \sum_{n=0}^{N-1} x(n)w(n)e^{-j2\pi\left(\frac{f}{f_s}\right)n}$$

If we sample the DFT at uniformly spaced points in the frequency domain:

$$X(k) = X(f)\Big|_{f=\frac{k}{N}f_s} = \sum_{n=0}^{N-1} x(n)w(n)e^{-j2\pi kn/N}$$

The spectrogram shown above was computed by sampling the speech signal at 10 kHz, applying a preemphasis filter ($H(z) = 1 - z^{-1}$), and weighting the signal with a 6 msec Hamming window. The net result is regarded as a wideband spectrogram (why?).

What happens if we increase the window duration to 30 msec?

If we constrain the DFT to a length that is a power of 2 (for example, 64 points were used above), and make use of the symmetry properties of the complex exponential function, we can dramatically decrease the number of computations (N log N complexity). This algorithm is known as a Fast Fourier Transform (FFT).

Time-Domain Windowing

Let $\{x(n)\}$ denote a sequence to be analyzed. Let's limit the duration of $\{x(n)\}$ to L samples:

$$\hat{x}(n) = x(n)w(n)$$

where $w(n)$ is a rectangular window and is defined as

$$w(n) = \begin{cases} 1, & 0 \le n \le L-1 \\ 0, & otherwise \end{cases}.$$

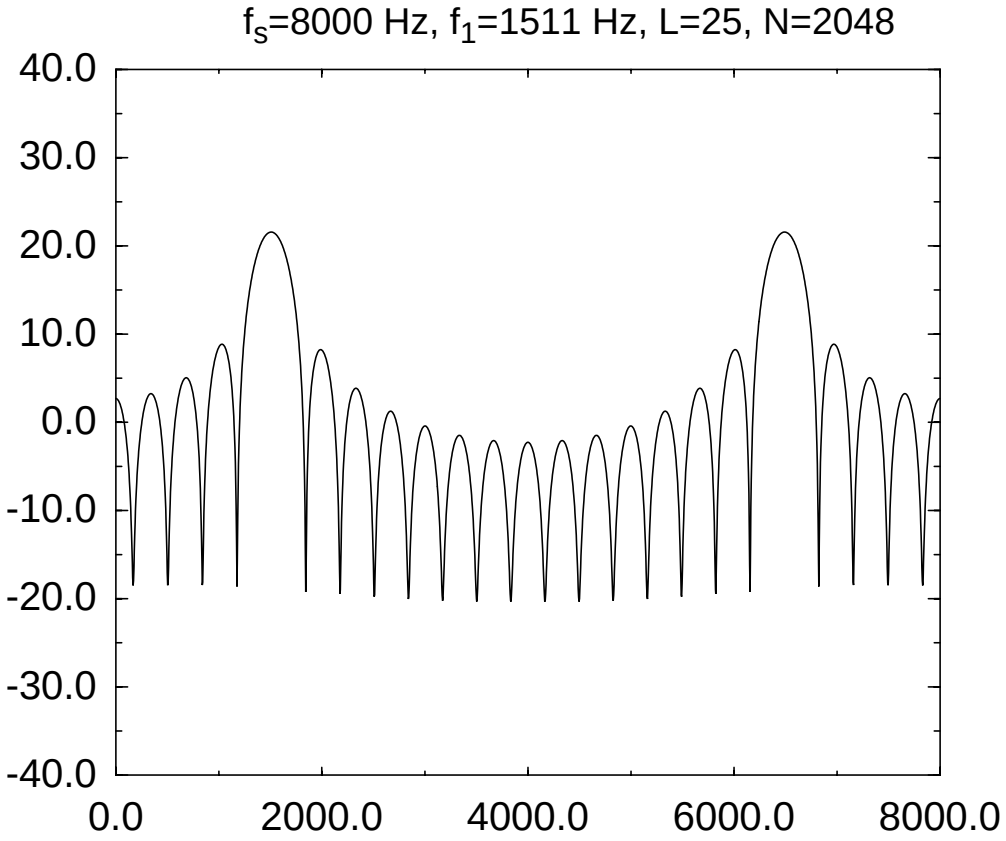
The Fourier transform of $w(n)$ is given by:

$$W(\omega) = \frac{\sin(\omega(L/2))}{\sin(\omega/2)} e^{-j\omega((L-1)/2)}$$

The transform of $\hat{x}(n)$ is given by:

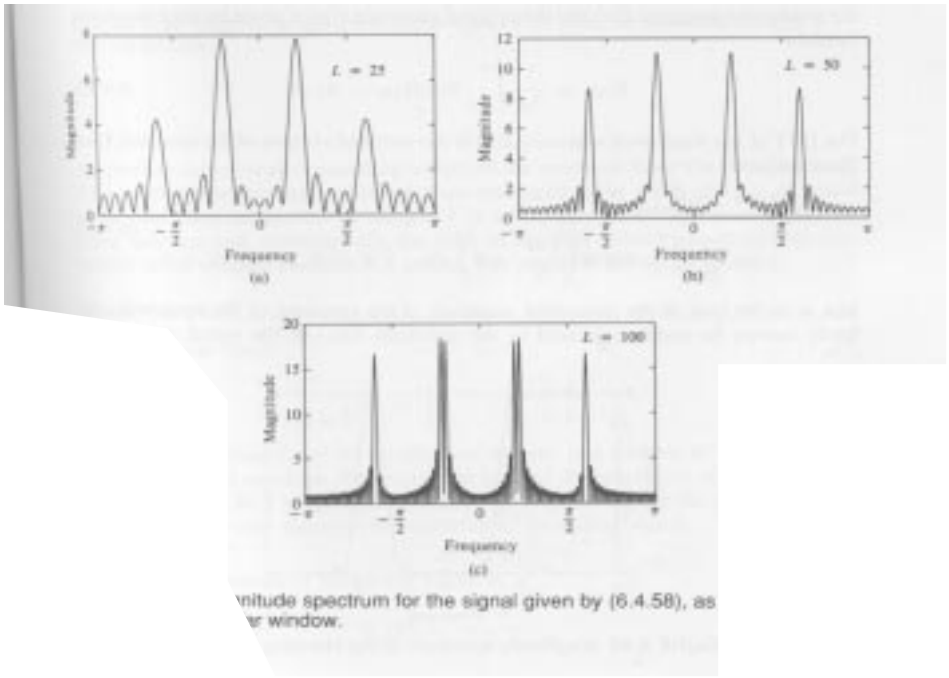
$$\hat{X}(\omega) = \frac{1}{2}[W(\omega - \omega_o) + W(\omega + \omega_o)].$$

This introduces frequency domain aliasing (the so-called picket fence effect):



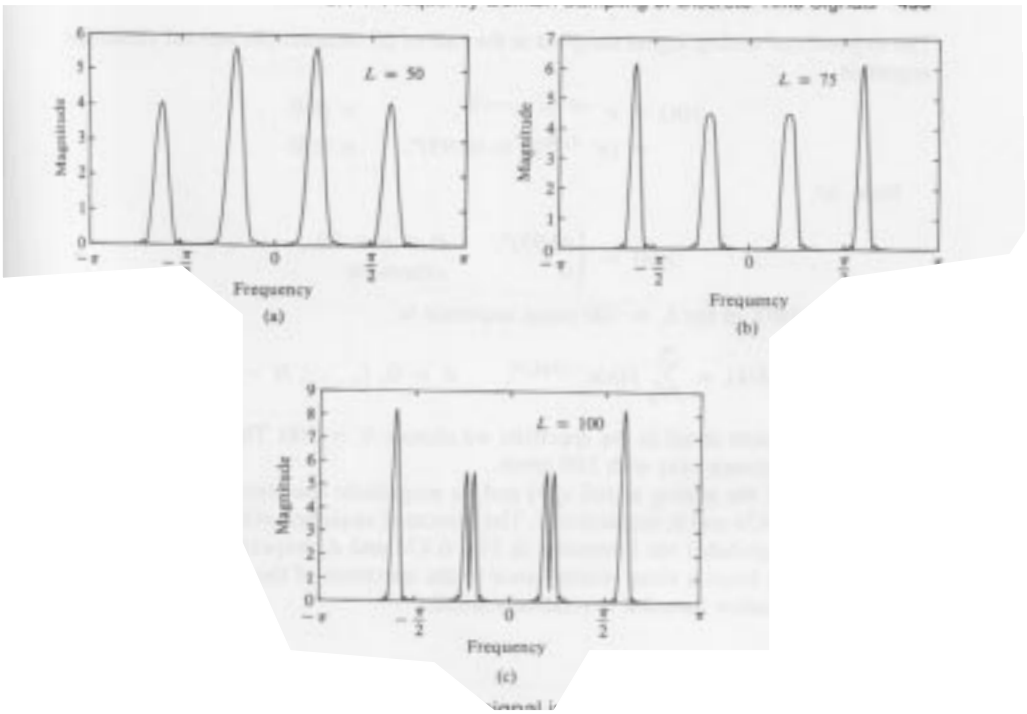
Improvements Via Better Windows

Rectangular Window:



nitude spectrum for the signal given by (6.4.58), as
r window.

Hanning Window:



Popular Windows

1. Rectangular:

$$w(k) = \begin{cases} 1, & |k| \le N \\ 0, & otherwise \end{cases}$$

2. Generalized Hanning:

$$w_H(k) = w(k)\left[\alpha + (1-\alpha)\cos\left(\frac{2\pi}{N}k\right)\right] \qquad 0 < \alpha < 1$$

$$\alpha = 0.54, \quad Hamming\ window$$

$$\alpha = 0.50, \quad Hanning\ window$$

3. Bartlett

$$w_B(k) = w(k)\left[1 - \frac{|k|}{N+1}\right]$$

4. Kaiser

$$w_K(k) = w(k) I_0\left(\alpha\sqrt{1 - \frac{K^2}{N}}\right) / I_0(\alpha)$$

5. Chebyshev:

$$w_N(k) = 2(x_0^2 - 1) w_{N-1}(k) + x_0^2[w_{N-1}(k-1) + w_{N-1}(k+1)] - w_{N-2}(k)$$

6. Gaussian

$$w_G(k) = \begin{cases} \exp\left[-\frac{1}{2}k^2 \tan^2\left(\frac{\theta_0}{2}\right)\right] & |k| < N \\ w_G(N-1) / \left[2N \sin^2\left(\frac{\theta_0}{2}\right)\right] & |k| < N \\ 0 & |k| > N \end{cases}$$

There are many others. The most important characteristics are the width of the main lobe and the attenuation in the stop-band (height of highest sidelobe). The Hamming window is used quite extensively.