Acoustics of Speech and Linear Prediction Modeling

EE6641 Analysis and Synthesis of Audio Signals

Yi-Wen Liu

Updated Nov. 24, 2015

Agenda

- Acoustics
  - Impedance, reflectance, multi-tube modeling
- Linear prediction (LP)
  - Relation to acoustics
  - LP and spectral analysis
    - Least-square formulation
    - Formants: resonant peaks
  - LP and speech synthesis
  - LP and speech recognition

Acoustics: the impedance concept

- Canonical acoustic variables are
  - Acoustic pressure P(x,t)
  - Volume velocity U(x,t)
  - Z = P/U is the **characteristic impedance**

    (and, how about Q = PU?)

  - Based on continuity and Newton's law, $Z = \rho c/A$
    - ρ: Density of air
    - c: Speed of sound (~340 m/s or 1 ft/ms)
    - A: Cross-section area

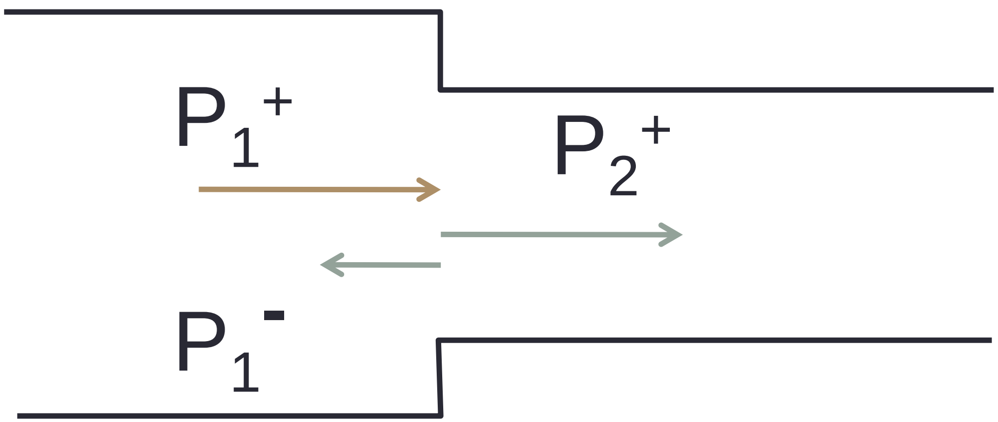
More on $Z = \rho c/A$

- The characteristic impedance is frequency-independent (?!)

- This formula does not consider
  - Dissipative loss
    - Friction 摩擦力
    - Viscosity 黏滯性
  - Propagation modes
  - Turbulence 紊流
  - Other nonlinear effects

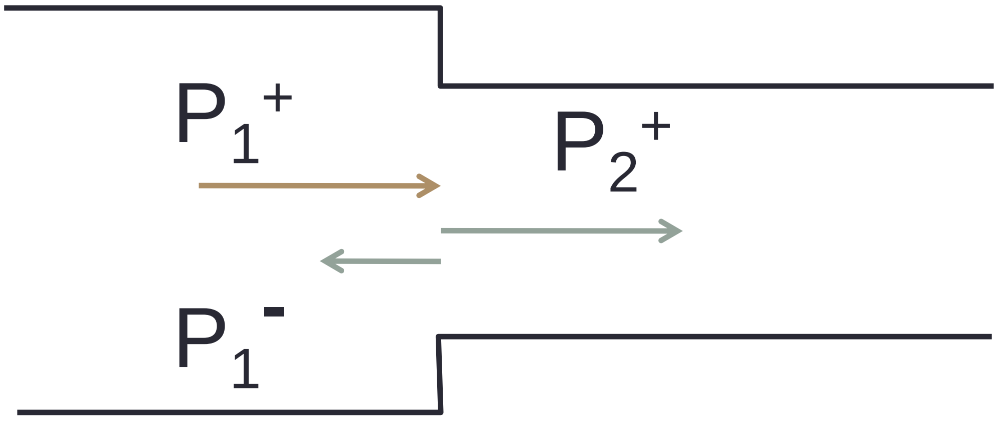
Continuity, impedance mismatch, and acoustic reflectance

- Now consider acoustic wave propagation through the boundary between two tubes
- Due to **impedance mismatch**, a portion of the wave is reflected and a portion crosses the boundary



Acoustic pressure reflectance and transmittance

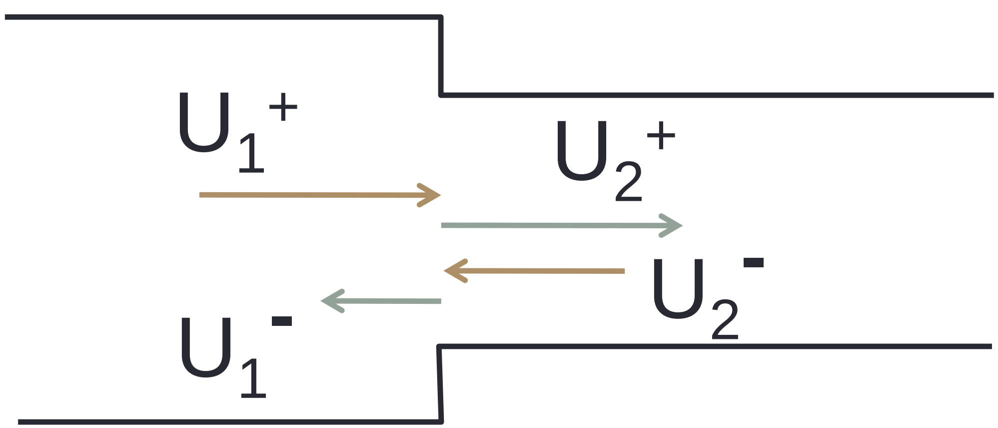
- Reflectance *R* can be defined as P_1^-/P_1^+
- Based on continuity, we can show that
  $r_{12} = (Z_1-Z_2)/(Z_1+Z_2)$
  - T = 1-r is called the transmittance



Two-port formulation

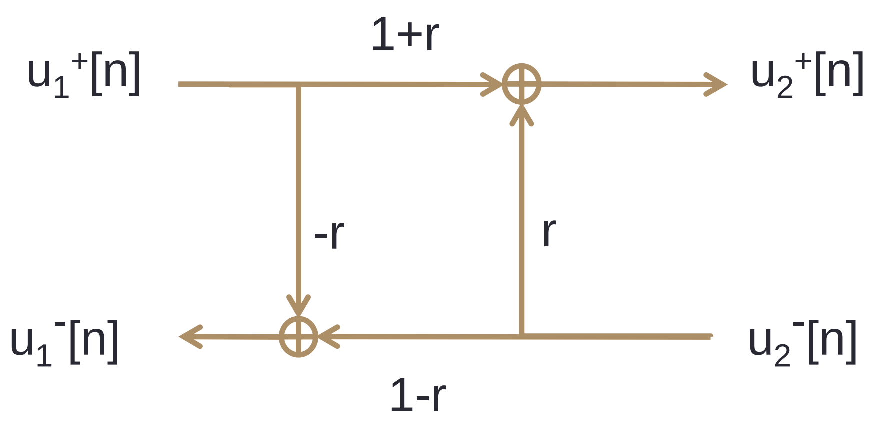
- Now consider reflected waves as a linear combination of left- and right- going incident waves
  - use volume velocity *U* as the input-output variable

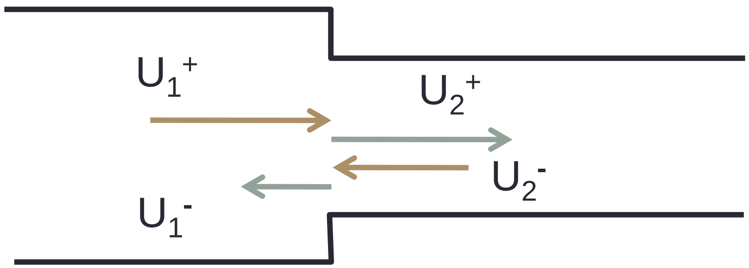
we have:
$$\begin{bmatrix} U_1^- \\ U_2^+ \end{bmatrix} = \begin{bmatrix} -r & 1-r \\ 1+r & r \end{bmatrix} \begin{bmatrix} U_1^+ \\ U_2^- \end{bmatrix}$$



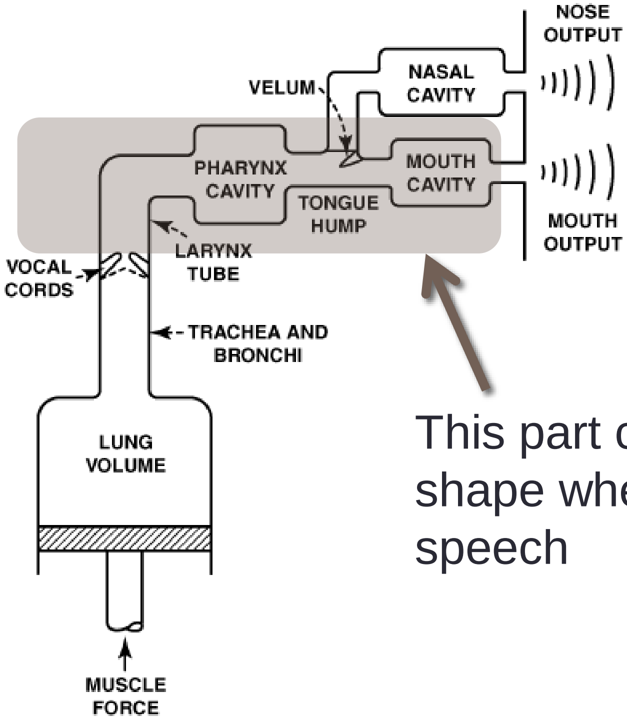
Signal-flow graph

$$\begin{bmatrix} U_1^- \\ U_2^+ \end{bmatrix} = \begin{bmatrix} -r & 1-r \\ 1+r & r \end{bmatrix} \begin{bmatrix} U_1^+ \\ U_2^- \end{bmatrix}$$



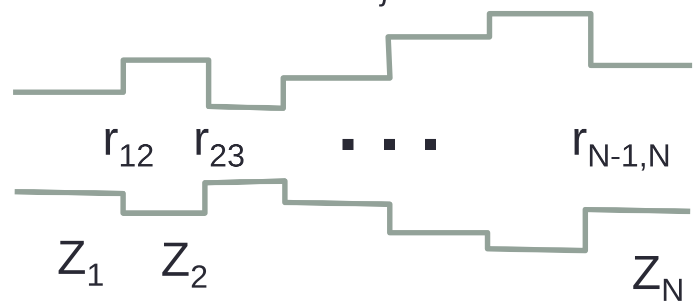


The multi-tube model of the vocal tract



This part changes its shape when producing speech

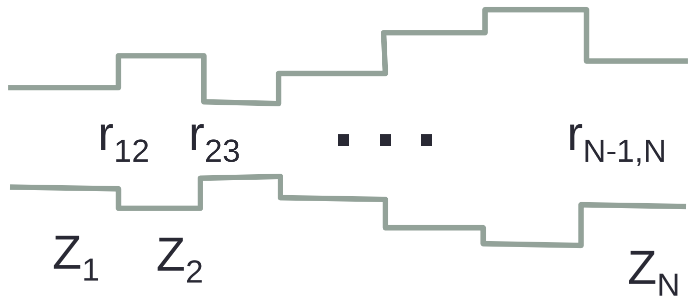
r_{ij} : reflectances

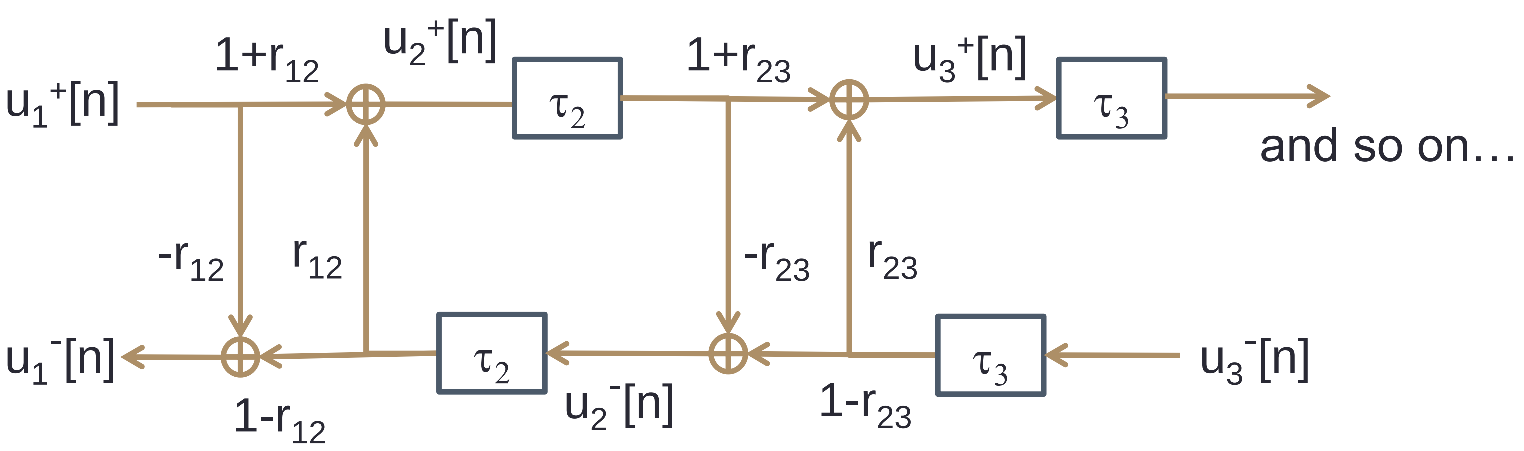


Z_i: char. impedances

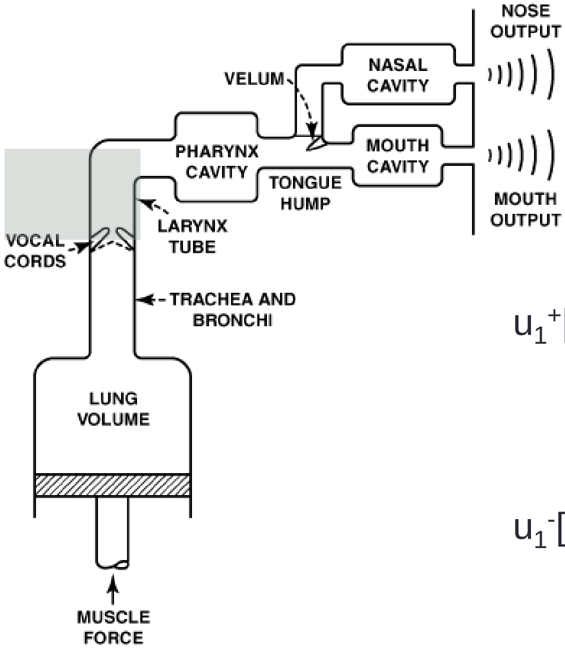
http://cobweb.ecn.purdue.edu/~ee649/notes/figures/vocal_schematic.gif

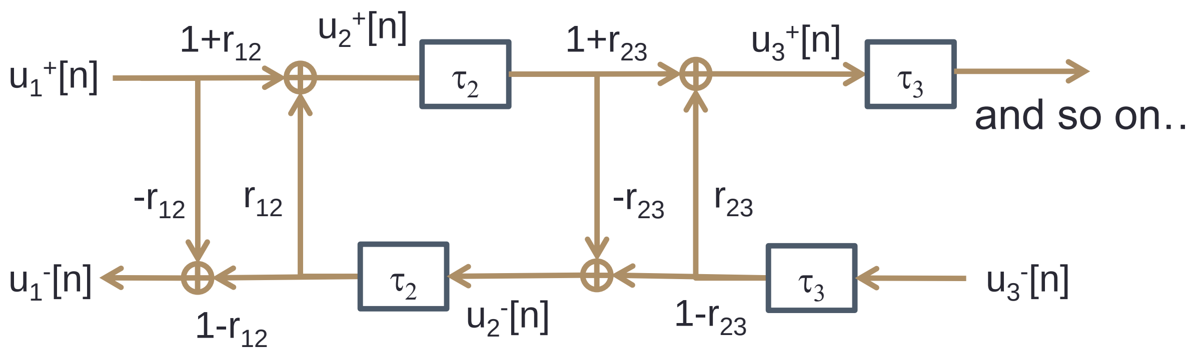
Signal flow of the multi-tube model





Source modeling



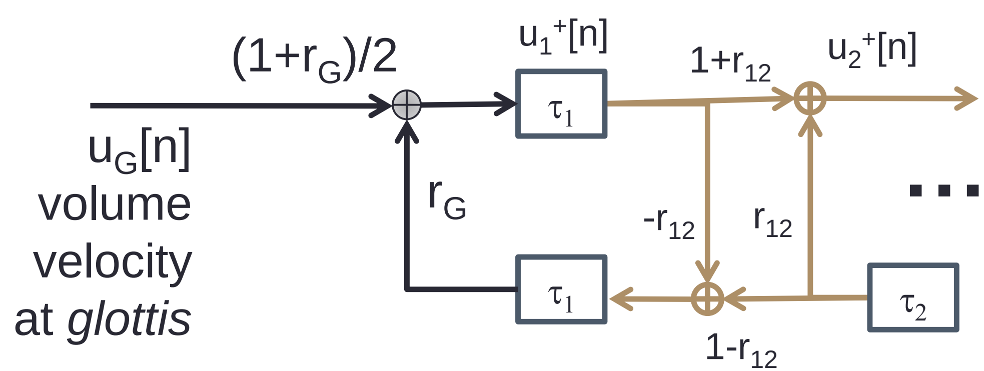


Source impedance:

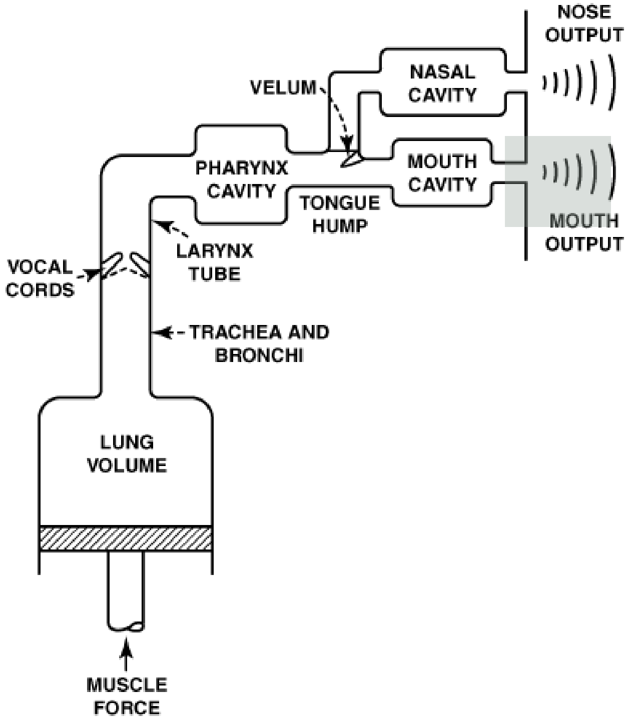
$$Z_G = R_G + j\omega L_G$$

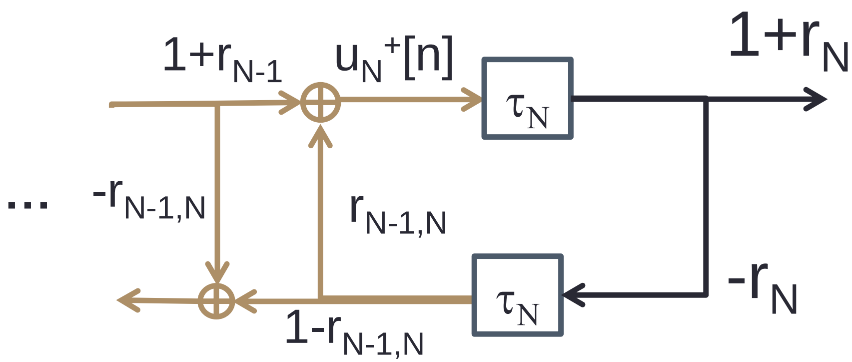
Source Reflectance:

$$r_G = \frac{Z_G - \frac{\rho c}{A_1}}{Z_G + \frac{\rho c}{A_1}}$$



Load modeling





Load impedance due to radiation from lips:

$$Z_r = R_r \parallel j\omega L_r$$

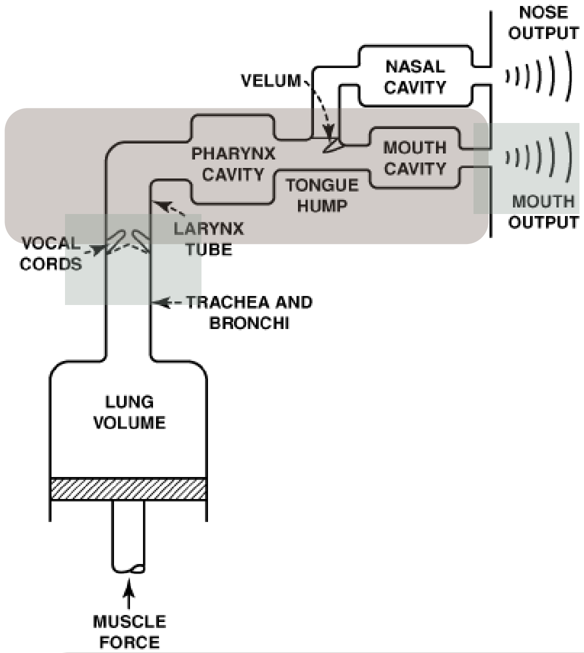
Load reflectance:

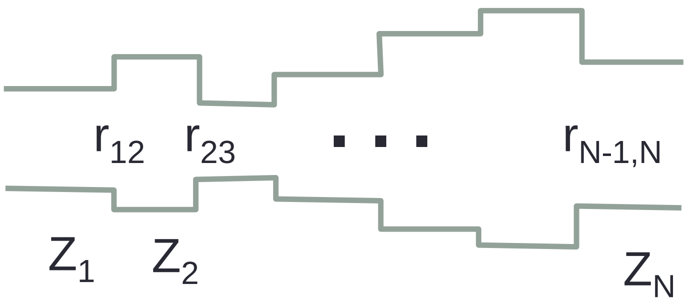
$$r_N = \frac{Z_N - Z_r}{Z_N + Z_r}$$

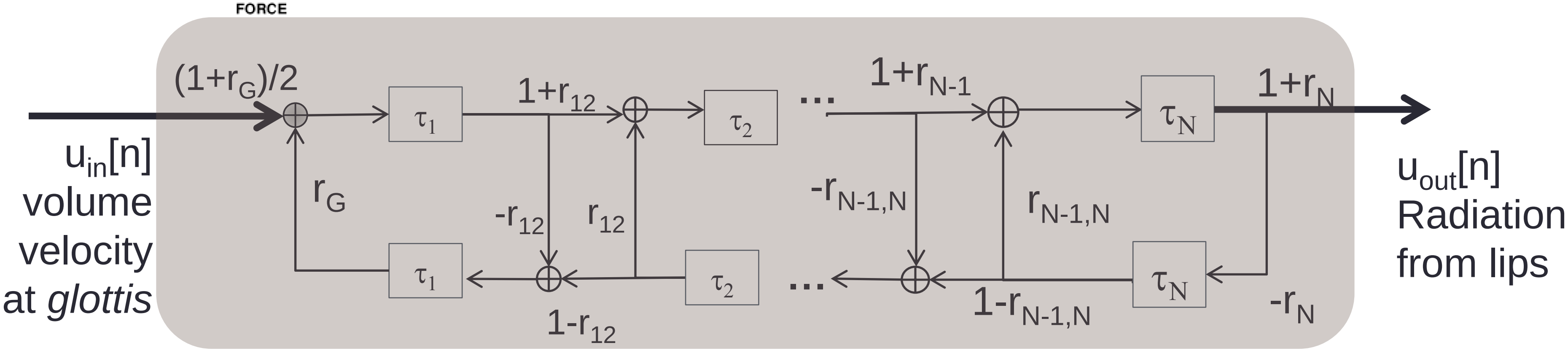
Frequency-dependence analysis:
Frequency decrease
=> Z_r approaches 0
=> r_N approaches 1,
which means output acoustic pressure approaches 0.

Frequency increases
=> r_N approaches a constant.

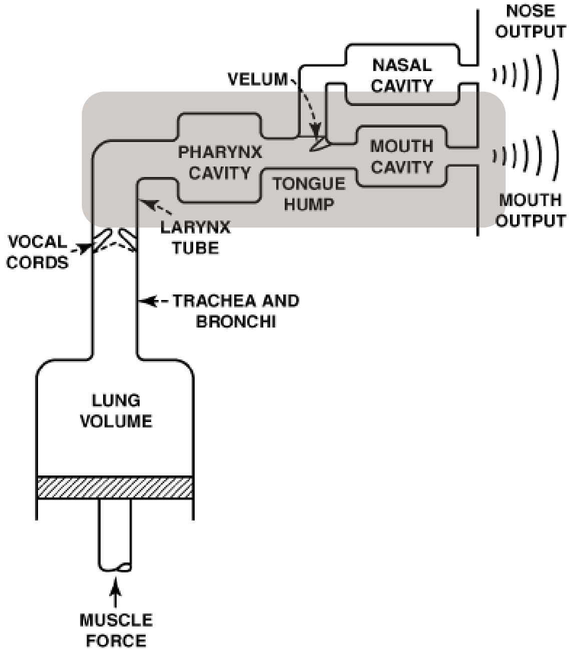
The entire multi-tube model with source and load

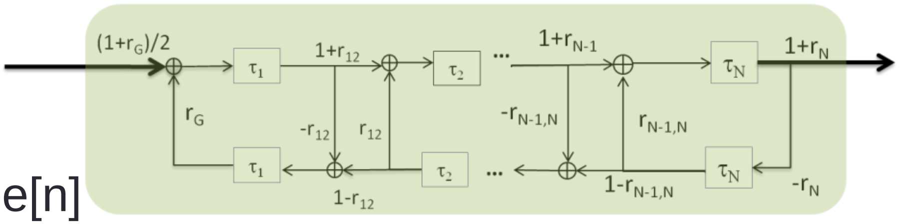






Linear prediction: Vocal tract as an all-pole IIR filter





e[n]
glottal source;
excitation

x[n]
output

We can write x[n] in terms of e[n]:

$$x[n] = \sum_{k=1}^{P} a_k x[n-k] + \Theta_0 e[n]$$

- This equation is called a **linear prediction model**
- *P* is the order of prediction
- a_k's are the linear prediction coefficients

Speech analysis and synthesis based on LP

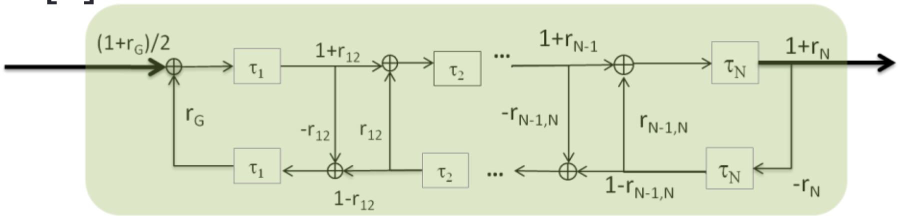
- Analysis:
  - record *x*[*n*]
  - estimate a_k's that gives best prediction of *x*[*n*]
  - Best prediction is formulated as a *least-square problem**

$$x[n] = \sum_{k=1}^{P} a_k x[n-k] + \Theta_0 e[n]$$

- Synthesis: Create *e*[*n*], synthesize *x*[*n*] in real-time.
  - Reflectances and LP coefficients are related via the *Levinson-Durbin recursive formula*.

glottal source; excitation e[n]

output x[n]



LP analysis: a least-square formulation

Find $\{a_k\}$, k = 1, 2, …, P

so as to minimize the sum of square of **prediction error** e[n], defined as below,

$$e[n] = x[n] - \sum_{k=1}^{P} a_k x[n-k]$$

Can be formulated as a matrix inverse problem:

$$\overset{\mathbf{K}}{\begin{bmatrix} x[1] & x[2] & \dots & \dots & x[P] \\ x[2] & x[3] & \dots & \dots & x[P+1] \\ & & \vdots & & \\ x[L] & x[L+1] & \dots & & x[P+L] \end{bmatrix}} \overset{\mathbf{a}}{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix}} \approx \overset{\mathbf{b}}{\begin{bmatrix} x[P+1] \\ x[P+2] \\ \vdots \\ x[P+L] \end{bmatrix}}$$

Solution:

$$\mathbf{a} = (\mathbf{K}^T\mathbf{K})^{-1}(\mathbf{K}^T\mathbf{b})$$

A least-square solver: MATLAB lpc() function

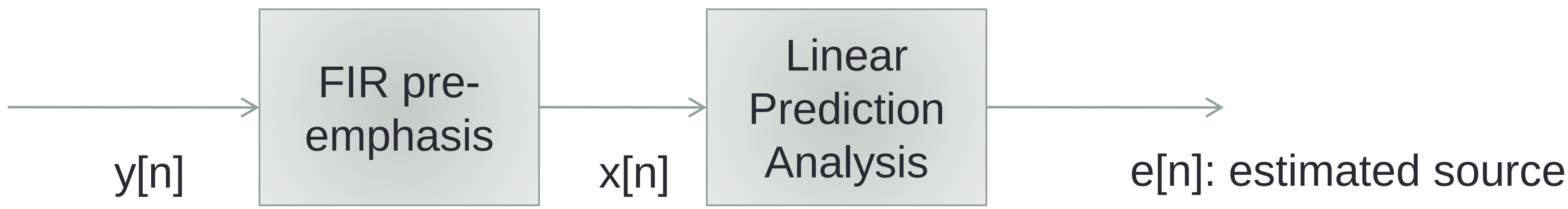
```
LPC  Linear Predictor Coefficients.
  A = LPC(X,N) finds the coefficients,
 A=[ 1 A(2) ... A(N+1) ], of an Nth order forward linear
 predictor.

      Xp(n) = -A(2)*X(n-1) - A(3)*X(n-2) - ... -
                   A(N+1)*X(n-N)

    such that the sum of the squares of the errors

      err(n) = X(n) - Xp(n)

    is minimized.
```

Remarks on least-square prediction

- The resulting prediction error $e[n]$ is *spectrally maximally flat*
  - The prediction "whitens" the signal
  - Makes sense, for the white noise is uncorrelated from sample to sample, which makes it impossible to predict further.

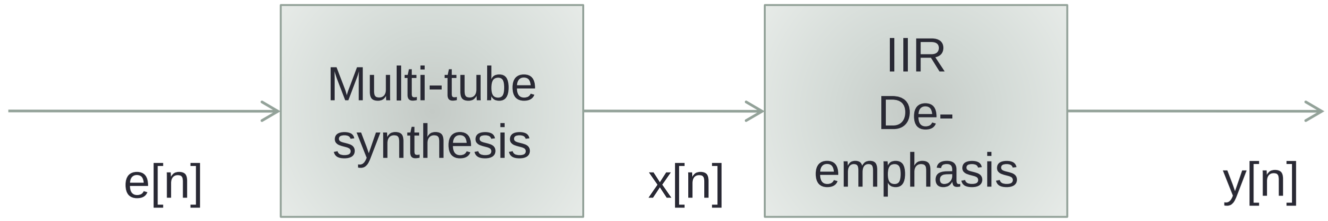
- In practice, because of spectral roll-off, one needs to *pre-emphasize** before LP analysis.

Pre-emphasis and de-emphasis



find $\{a_1, \ldots a_P\}$

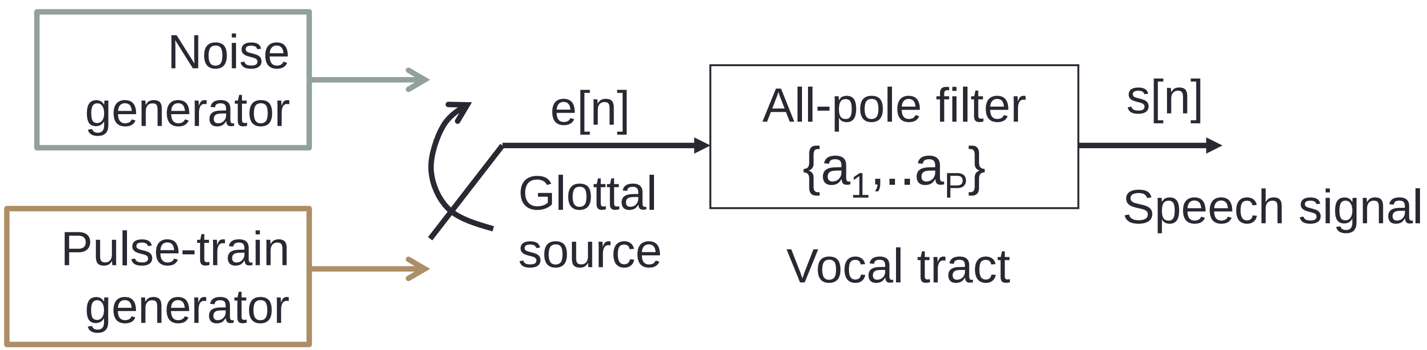
$x[n] = y[n] - 0.95\, y[n-1]$



$y[n] = 0.95\, y[n-1] + x[n]$

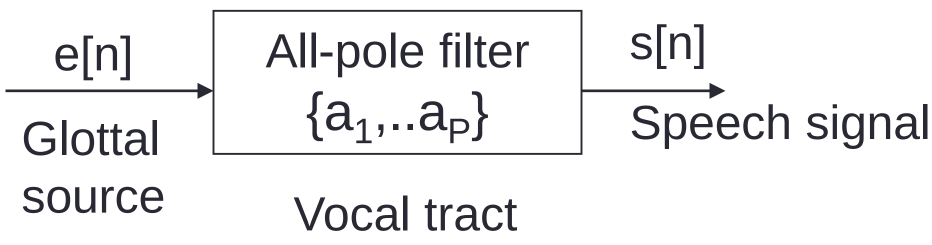
Source-filter separation

- Find $\{a_1,..a_P\}$ such that energy of e[n] is minimized.
  - Turns out that such e[n] will be maximally spectrally flat.
- This provides a **source-filter separation**:
  - $\{a_1,..a_P\}$: vocal-tract filter
  - e[n]: glottal source = {voiced, unvoiced}
    - When voiced, use pulse train
    - When unvoiced, use white noise

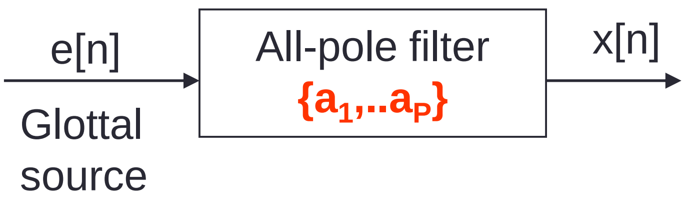


More on LP

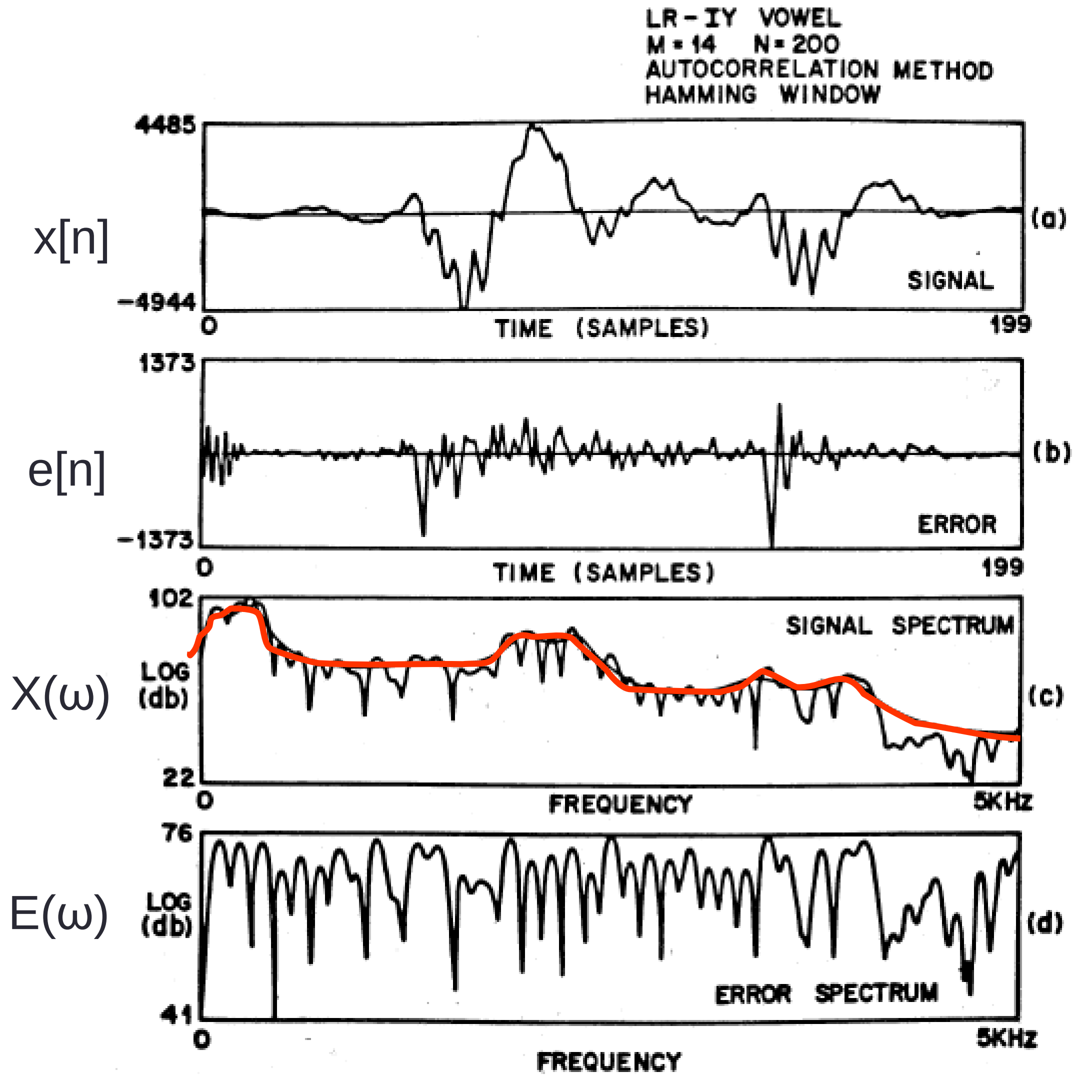
- **Speech synthesis**: By replacing e[n] with a template, speech compression achieves <8k bits/s.
  - Codebook excited linear prediction (CELP)
  - key technology for voice over internet and wireless networks.
- **Speech recognition**: From $\{a_1,..a_P\}$, we can estimate
  - Vocal tract constriction
  - Frequency-envelope; formant structure.



LP finds an all-pole filter that provides spectral smoothing



$$F(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}}$$

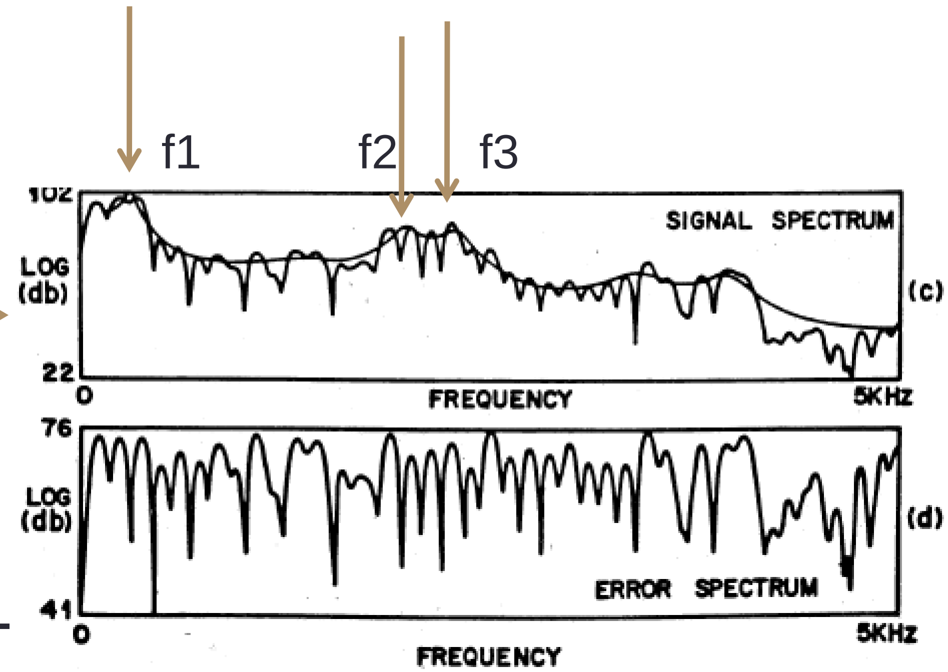


Adapted from Rabiner & Juang, p. 108.

Formant frequencies and speech production

$$F(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}}$$

The formant filter



2000 **f2-f1** 300

300

i (heed)

u (loose)

ɪ (hid)

ʊ (look)

f1

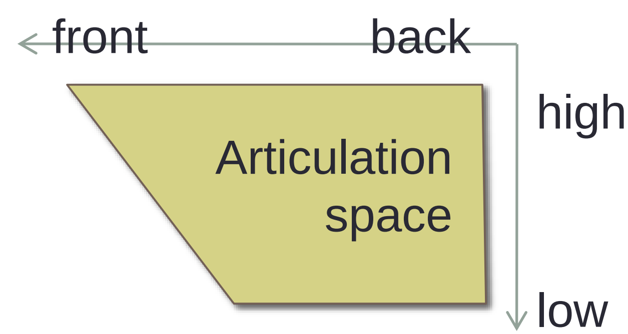
ɛ (head)

ɔ (law)

(had)

ɑ (lot)

900 Hz



References

- 王小川：語音訊號處理(三版，2009）全華出版社
- S. Vaseghi (2007). Multimedia signal processing: theory and applications in speech, music, and communications. London: Wiley.
- T. Quatieri (2001). Discrete-time speech signal processing: principles and practice. Prentice-Hall.
- L. Rabiner and B.-H. Juang. (1993). Fundamentals of speech recognition. Prentice-Hall.