6.300 Signal Processing

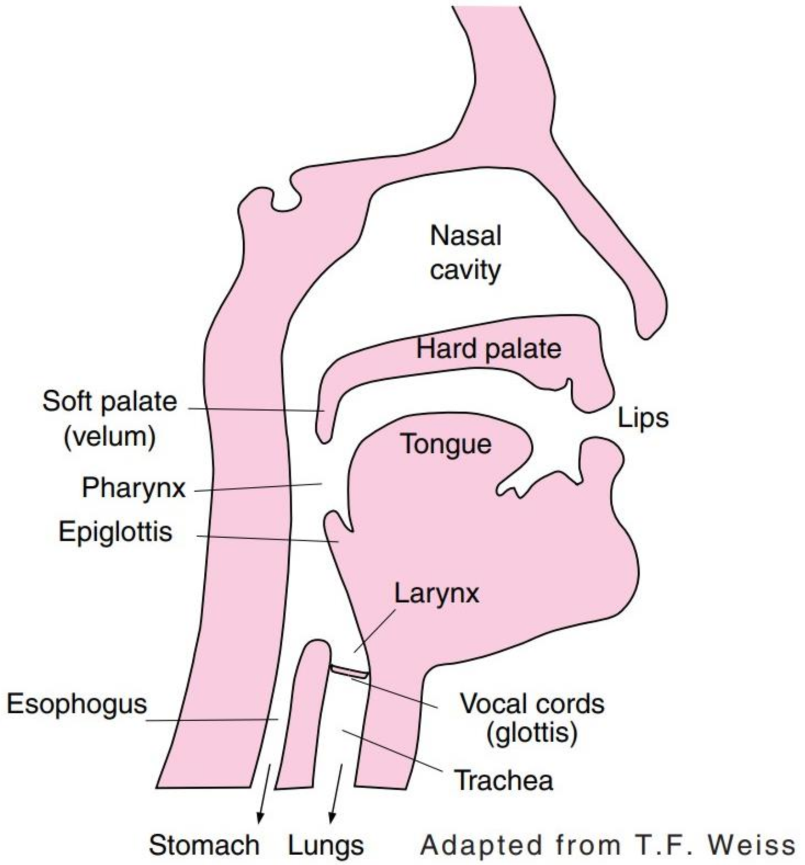
Week 11, Lecture B:
Speech

- Source/Filter Model of Speech Production
- Speech Analysis
- Speech Synthesis

Lecture slides are available on CATSOOP:
https://sigproc.mit.edu/fall24

Speech

Speech is generated by the passage of air from the lungs, through the vocal cords, mouth, and nasal cavity.



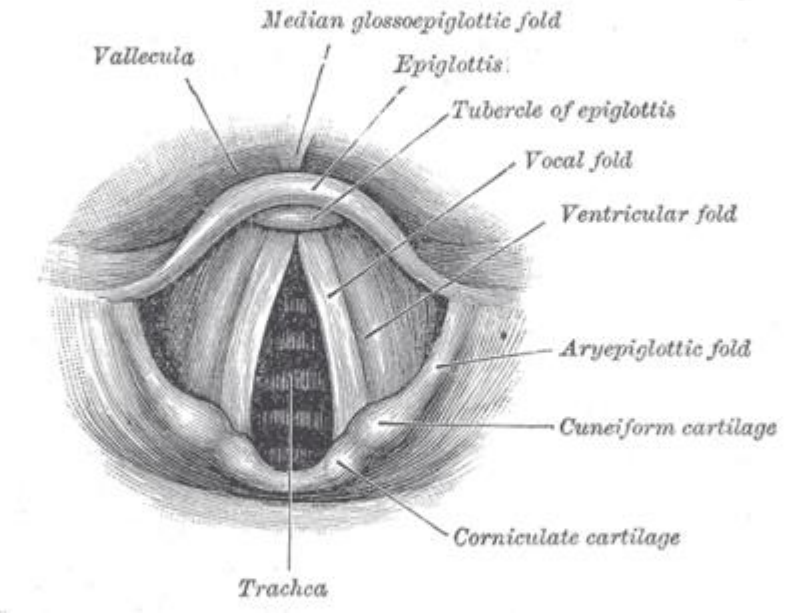
We can think of this as having two parts: the source and the filter.

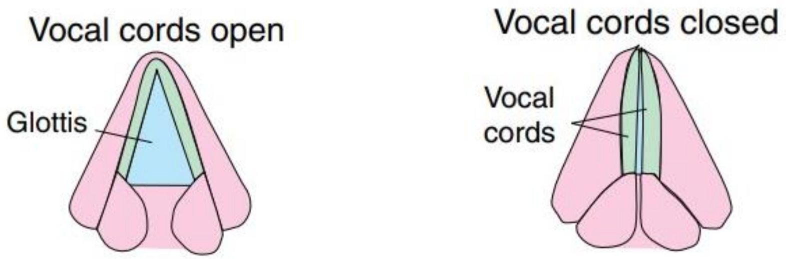
Source/Filter Model of Speech Production

Controlled by complicated muscles, vocal cords are set in vibration by the passage of air from the lungs.

Looking down the throat

Gary's Anatomy





Adapted from T.F. Weiss

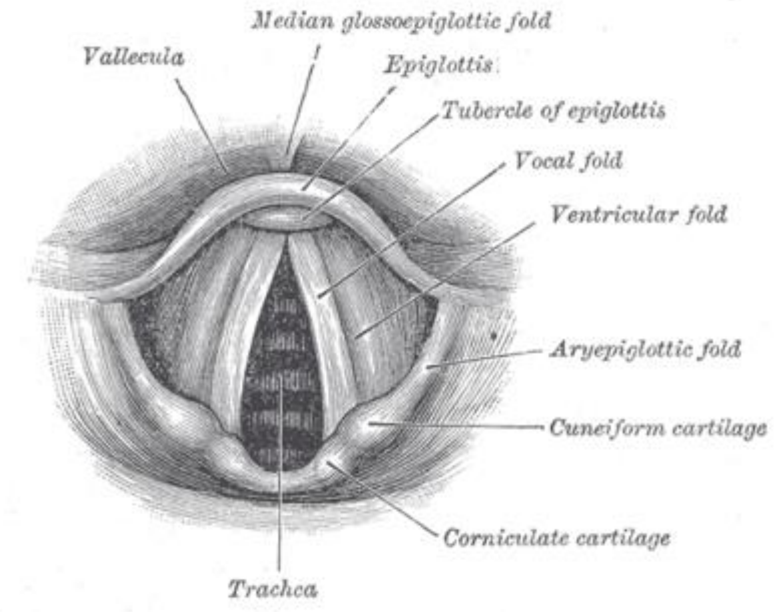
During voiced speech, the glottis generates puffs of air that are a few ms in duration. The frequency of puffs ranges from 100- 300 Hz.

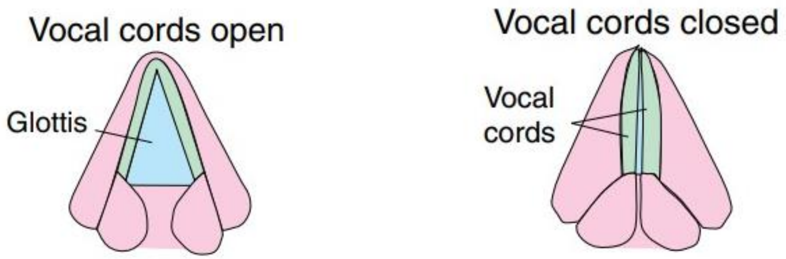
Source/Filter Model of Speech Production

Sound is produced when the air which passes through the vocal cords causes them to vibrate and create sound waves.

Looking down the throat

Gary's Anatomy





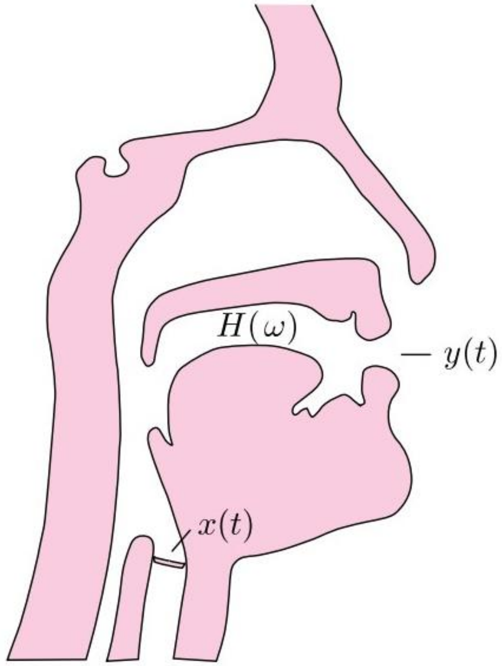
Adapted from T.F. Weiss

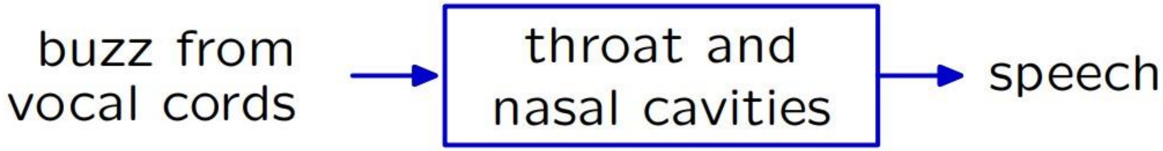
In male adults, the vocal folds are usually 17-23 mm long, and 12.5 -17 mm in female adults. They may be stretched 3 or 4 mm by action of the muscles in the larynx.

The male speaking voice averages about 125 Hz, while the female voice averages about 210 Hz. Children's voices average over 300 Hz.

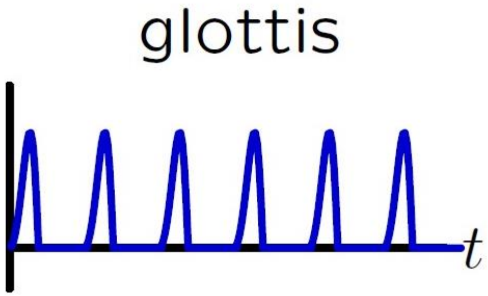
Source/Filter Model of Speech Production

Vibrations of the vocal cords are “filtered” by the mouth and nasal cavities to generate speech.



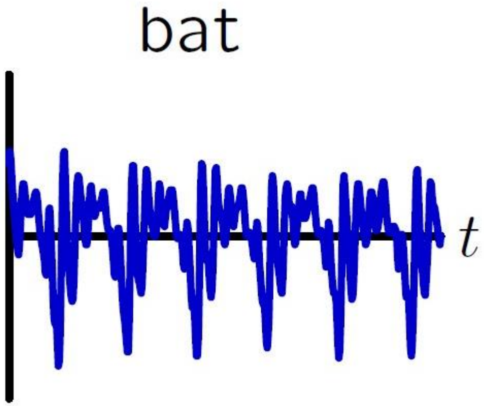


Source/Filter Model of Speech Production



Speech sound:

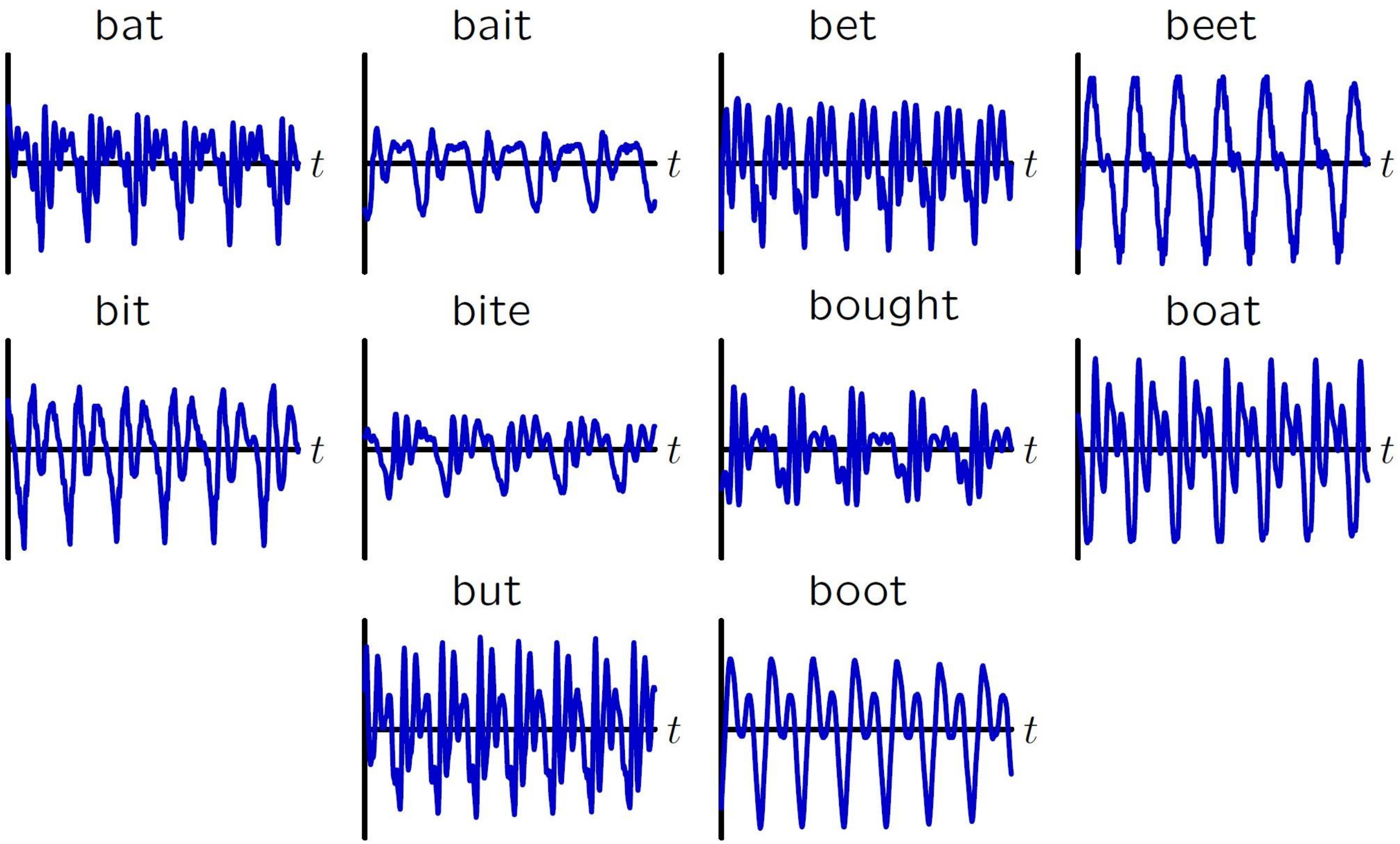
- Vowel
- Consonant



Vowel sounds are periodic, because the glottis vibrates periodically.

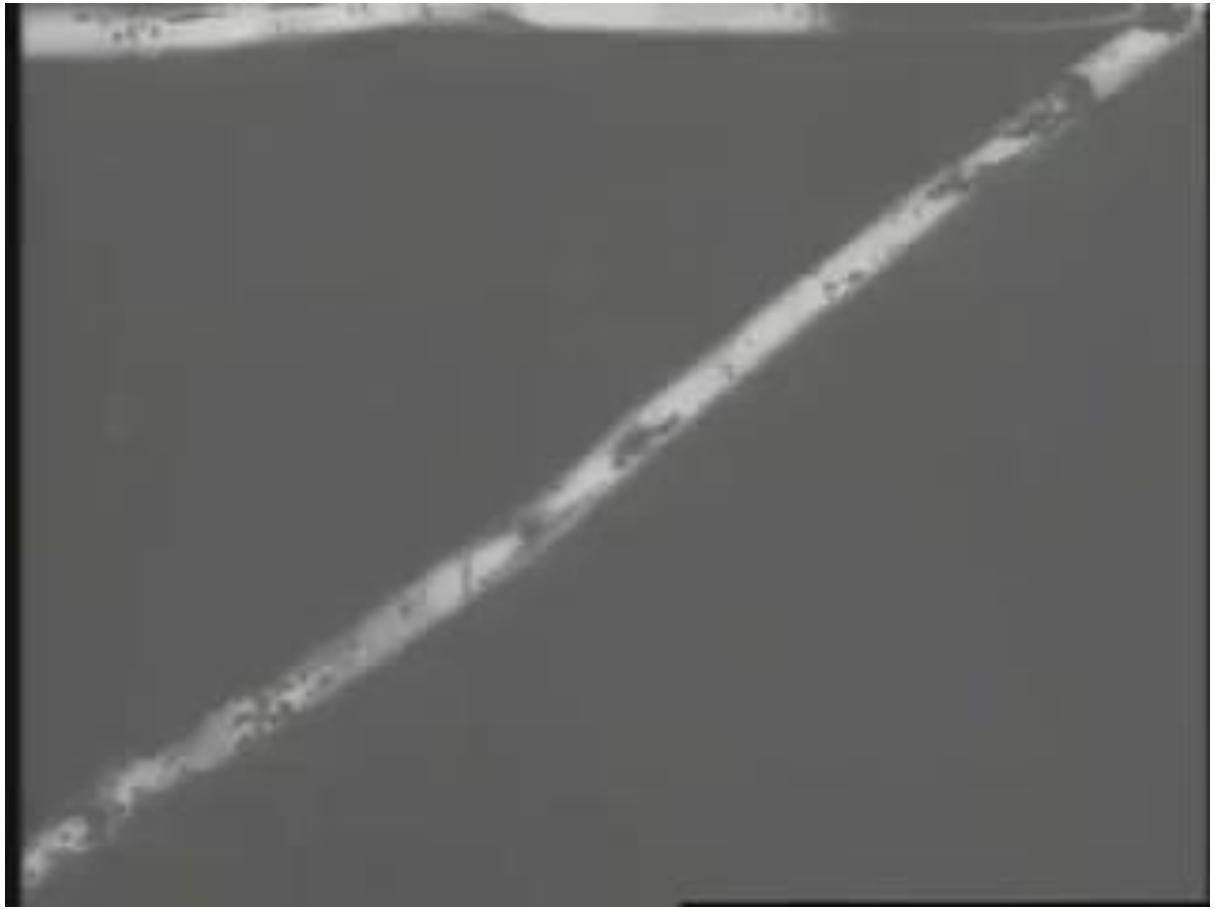
Source/Filter Model of Speech Production

Vowels sound different because mouth and lip positions are different.



Speech Production

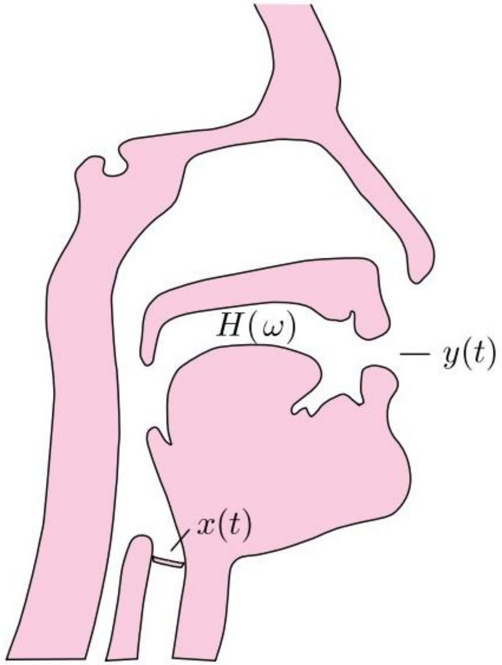
X-ray movie showing speech in production.



By Prof. Kenneth Noble Stevens (https://en.wikipedia.org/wiki/Kenneth_N._Stevens)

Demonstration

Physical model of the vocal tract.



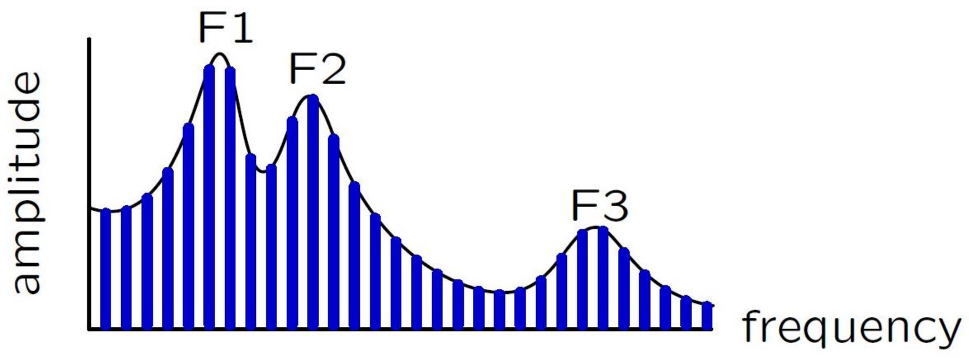
buzz from vocal cords → throat and nasal cavities → speech

Buzzer represents sound from glottis.
Machined cavities represent vocal tract.

Chiba and Kajiyama Model replicated by Takayuki Arai.

Formants

Resonant frequencies of the vocal tract.



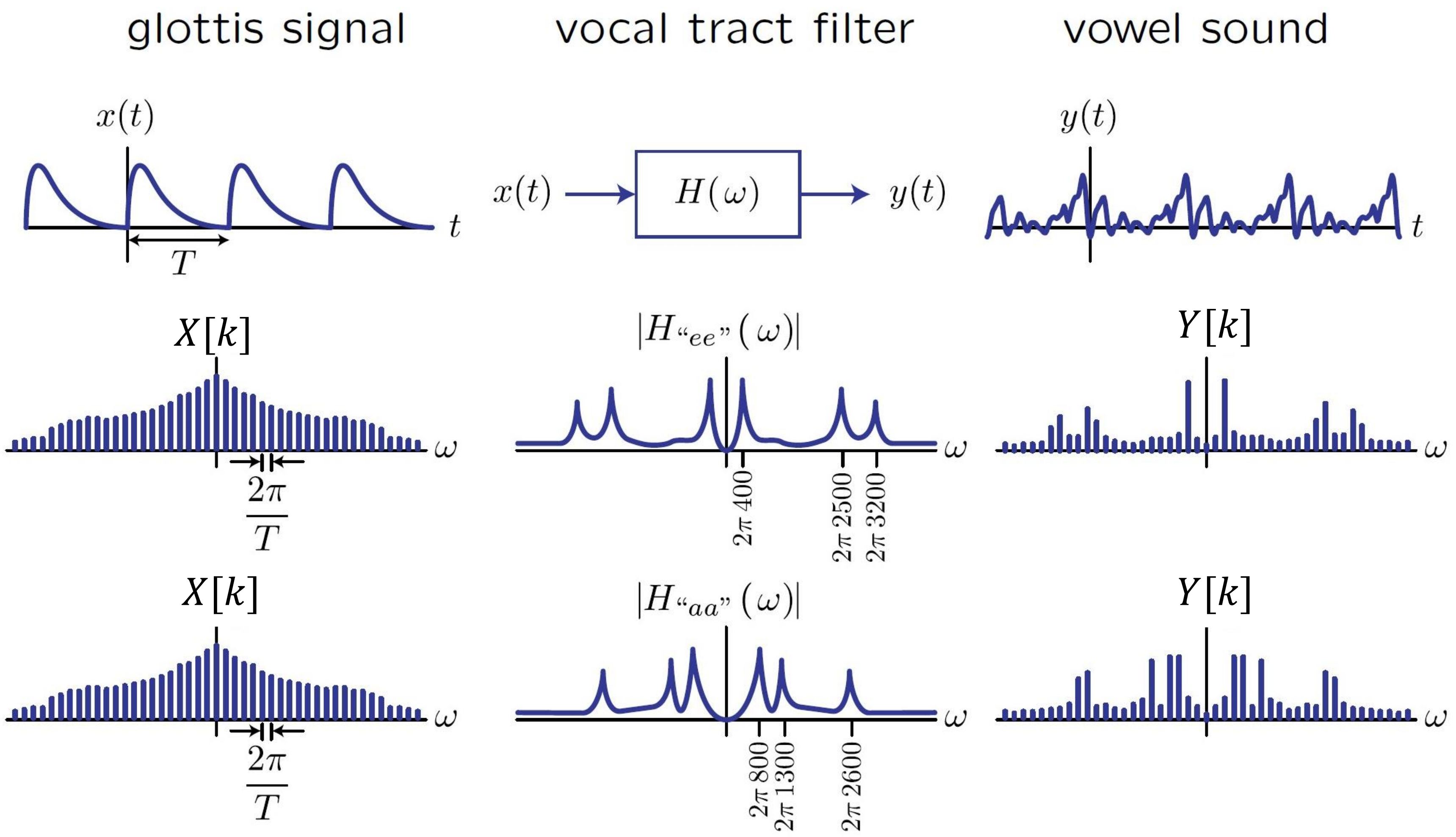
	Formant	heed	head	had	hod	haw'd	who'd
Men	F1	270	530	660	730	570	300
	F2	2290	1840	1720	1090	840	870
	F3	3010	2480	2410	2440	2410	2240
Women	F1	310	610	860	850	590	370
	F2	2790	2330	2050	1220	920	950
	F3	3310	2990	2850	2810	2710	2670
Children	F1	370	690	1010	1030	680	430
	F2	3200	2610	2320	1370	1060	1170
	F3	3730	3570	3320	3170	3180	3260

Average resonance frequencies of the first three formants (F1, F2, F3) of the vowels of men, women and children (from Appleton and Perera, eds., *The Development and Practice of Electronic Music*, Prentice-Hall, 1975, p.42; after Peterson and Barney, *Journal of the Acoustical Society of America*, vol. 24, 1952, pp. 175-84).

https://www.sfu.ca/sonic-studio-webdav/handbook/Formant.html

Speech Production

Same glottis signal + different formants -> different vowels.



We detect changes in the filter function to recognize vowels.

Singing

We detect changes in the filter function to recognize vowels ... at least sometimes.

Demonstration.

“la" scale.

“lore" scale.

“loo" scale.

“ler" scale.

“lee" scale.

Low Frequency: “la" “lore" “loo" “ler" “lee".

High Frequency: “la" “lore" “loo" “ler" “lee".

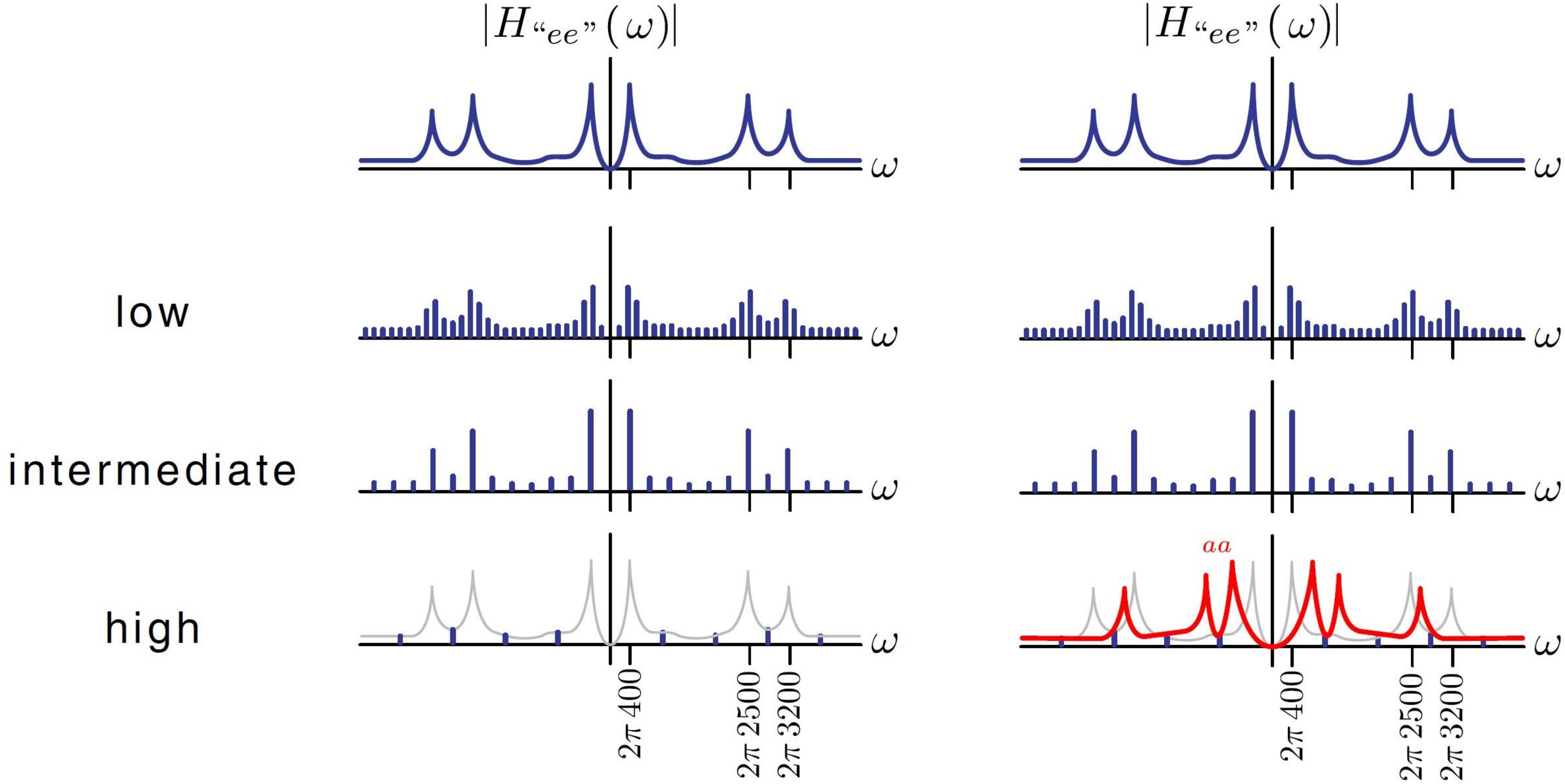
Why can’t we distinguish the vowels at higher frequency?

Participation question for Lecture

http://www.phys.unsw.edu.au/jw/soprane.html

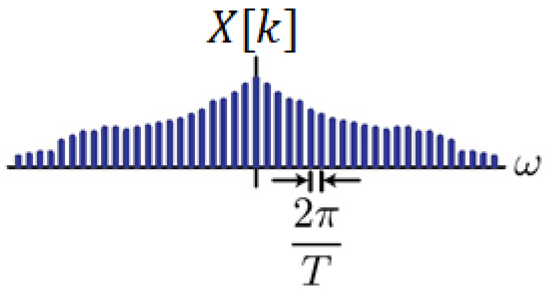
Speech Production

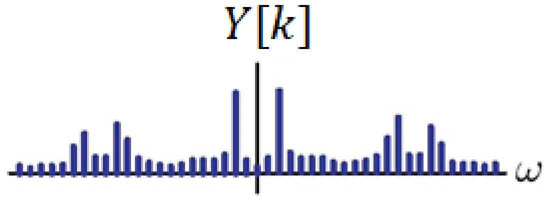
We detect changes in the filter function to recognize vowels.

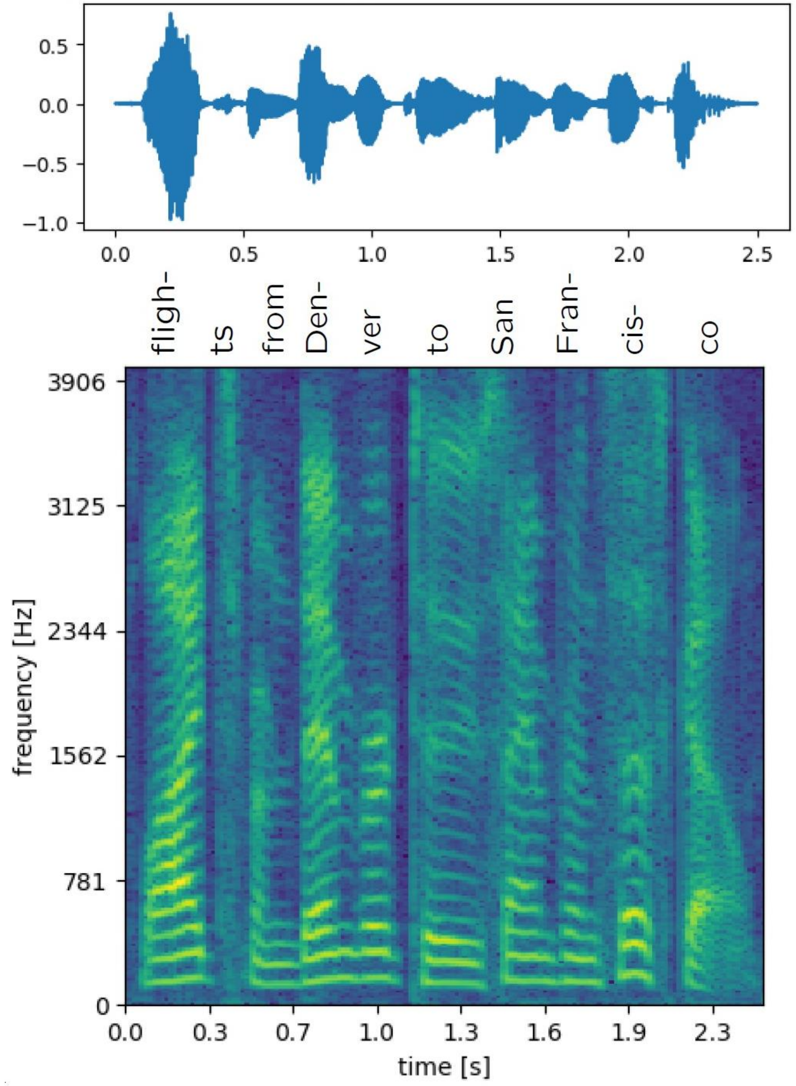


Time and Frequency Structure of Speech

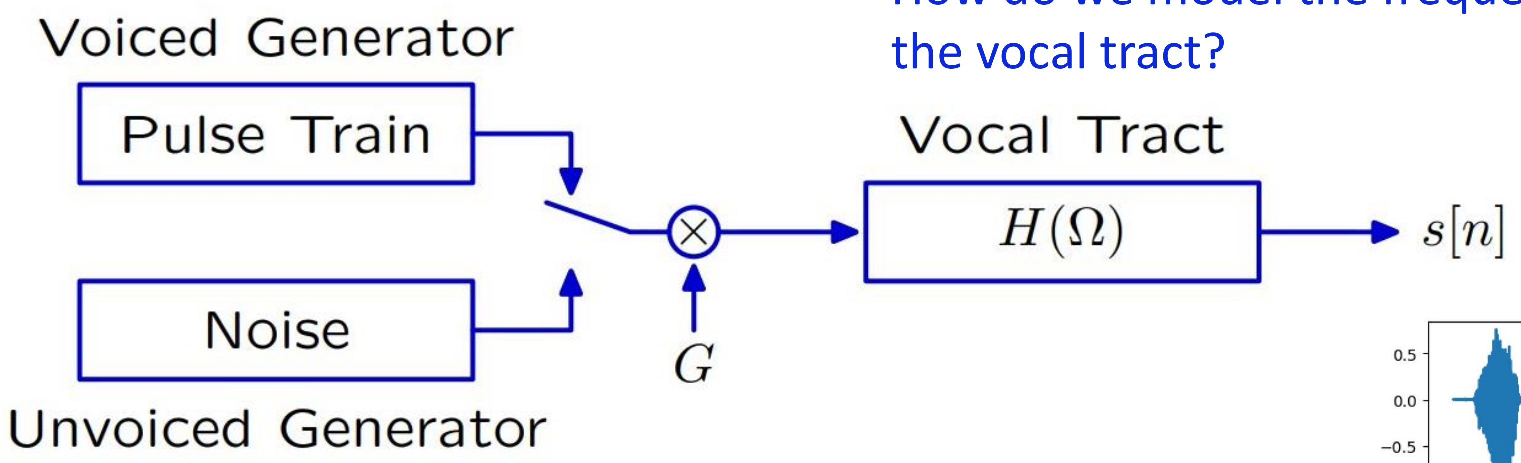
Time plot & spectrogram of “fights from Denver to San Francisco."



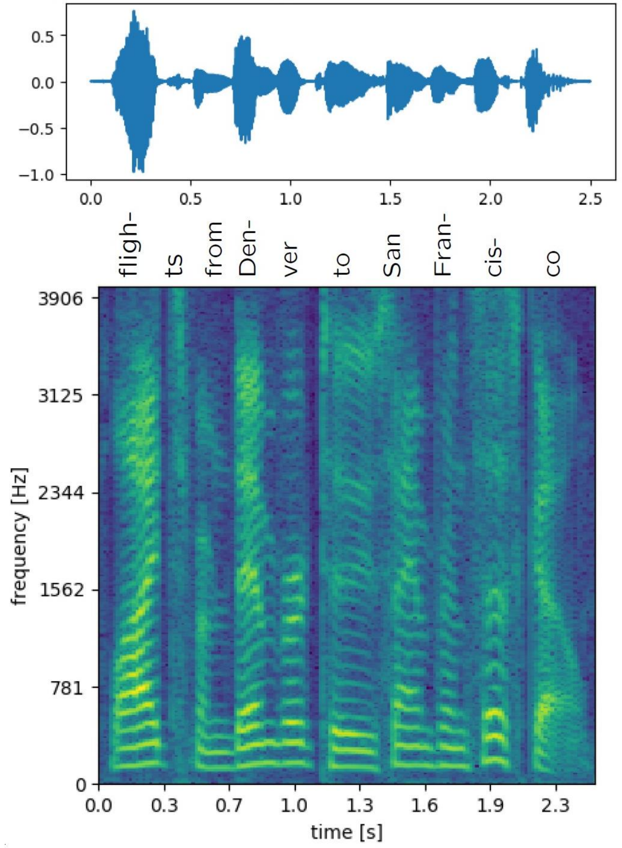




Source/Filter Model



How do we model the frequency response of the vocal tract?



Acoustic Sources:

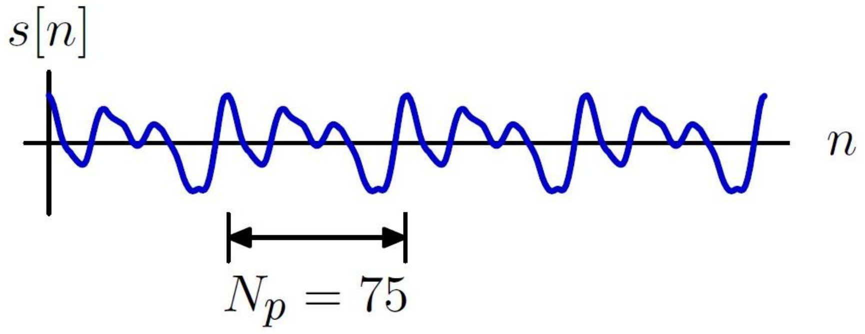
- Pulse train for voiced utterances
- Gaussian noise for unvoiced utterances

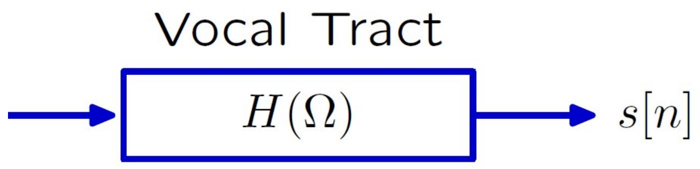
Gain: G controls loudness

Vocal Tract: filter represents shape of mouth, tongue, and lips

Linear Predictive Coding

The speech signal $s[n]$ is shaped by the vocal tract.





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

Develop a “predictive" model:

$$s[n] = \sum_{k=1}^{P} a_k s[n-k] \qquad h[n] = \delta[n] + \sum_{k=1}^{P} a_k h[n-k] \quad \text{for n>0}$$

where output at time *n* is a linear combination of *P* previous outputs.

Estimate a_k.

Linear Predictive Coding

Let $\hat{s}[n]$ represent our prediction for $s[n]$.

$$\hat{s}[n] = \sum_{k=1}^{P} a_k s[n-k]$$

And then minimize the squared difference between $s[n]$ and $\hat{s}[n]$:

$$E = \sum_{n=1}^{N_p} (s[n] - \hat{s}[n])^2 \quad = \sum_{n=1}^{N_p} \left(s[n] - \sum_{k=1}^{P} a_k s[n-k]\right)^2$$

Set the derivative of *E* with respect to a_i equal to zero for $1 \le i \le P$:

$$\frac{\partial E}{\partial a_i} = 0 \;= \sum_{n=1}^{N_p} 2\left(s[n] - \sum_{k=1}^{P} a_k s[n-k]\right) \cdot (-s[n-i]) \;= -2 \sum_{n=1}^{N_p} \left(s[n]s[n-i] - \sum_{k=1}^{P} a_k s[n-k] \cdot s[n-i]\right)$$

Therefore

$$\sum_{n=1}^{N_p} s[n]s[n-i] = \sum_{k=1}^{P} a_k \sum_{n=1}^{N_p} s[n-k] \cdot s[n-i] \qquad \text{for } 1 \le i \le P.$$

Linear Predictive Coding

$$\sum_{n=1}^{N_p} s[n]s[n-i] = \sum_{k=1}^{P} a_k \sum_{n=1}^{N_p} s[n-k] \cdot s[n-i]$$

For the above expression we can rewrite in terms of the autocorrelation function:

$$R[i] = \sum_{n=1}^{N_p} s[n]s[n-i]\,, \qquad R[i] = R[-i]$$

Our final result is

$$R[i] = \sum_{k=1}^{P} a_k R[i-k] \quad \text{for } 1 \le i \le P.$$

This can be written as a matrix equation:

$$\begin{bmatrix} R[0] & R[1] & \cdots & R[P-1] \\ R[1] & R[0] & \cdots & R[P-2] \\ R[2] & R[1] & \cdots & R[P-3] \\ \cdots & \cdots & \cdots & \cdots \\ R[P-1] & R[P-2] & \cdots & R[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_P \end{bmatrix} = \begin{bmatrix} R[1] \\ R[2] \\ R[3] \\ \cdots \\ R[P] \end{bmatrix}$$

Linear Predictive Coding

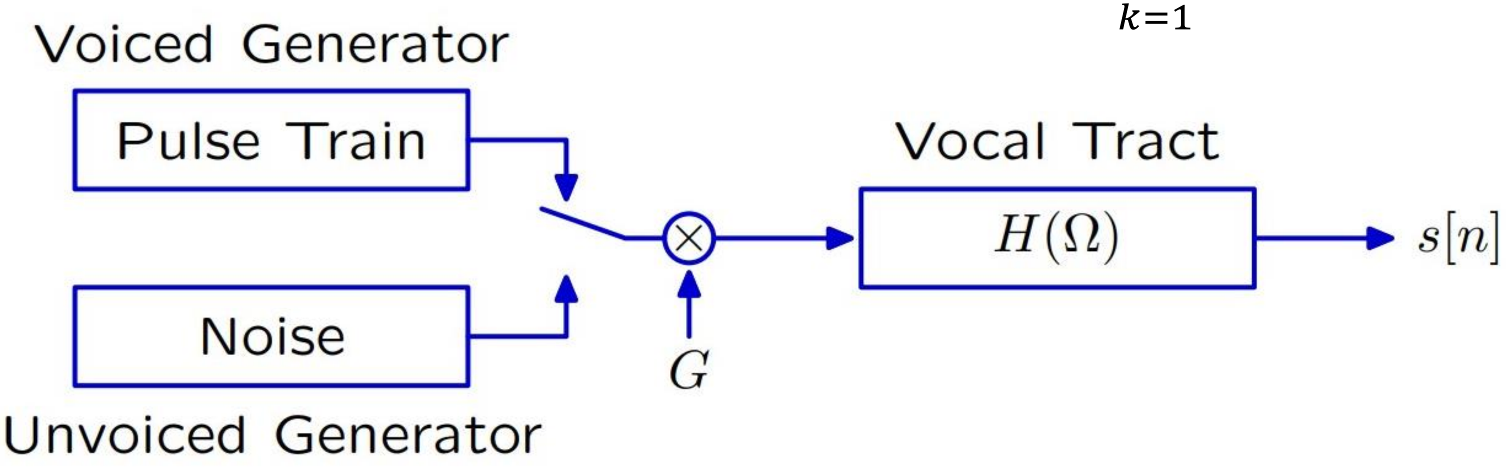
Summary of LPC procedure:

1. Take s[n], select a region of time using a window function w[n]
2. calculate the autocorrelation function R[i]
3. solve the set of linear equations to find a_k.

$$R[i] = \sum_{n=1}^{N_p} s[n]s[n-i]$$

Now $H(\Omega)$ is represented by the following difference equation:

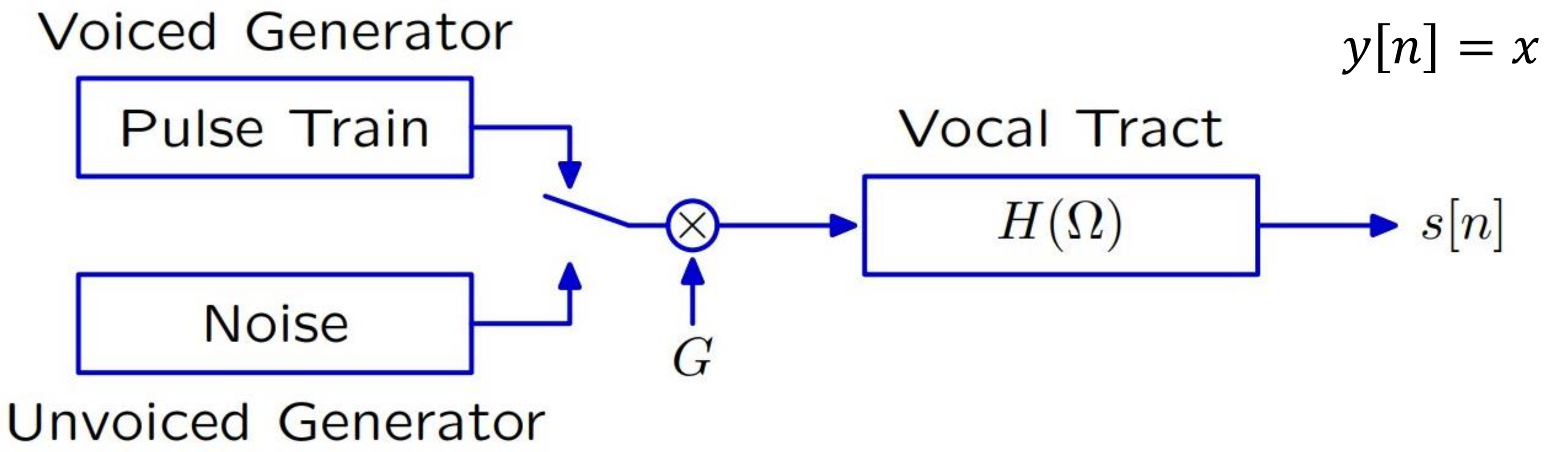
$$y[n] = x[n] + \sum_{k=1}^{P} a_k y[n-k]$$



Check yourself!

Given the difference equation, how would we find the frequency response of $H(\Omega)$?

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



Method 1:

Set $x[n] = \delta[n]$, use the difference equation to find $h[n]$, then find $H(\Omega)$

Check yourself!

Given the difference equation, how would we find the frequency response of $H(\Omega)$?

Method 2:

Take the Fourier transform of the difference equation:

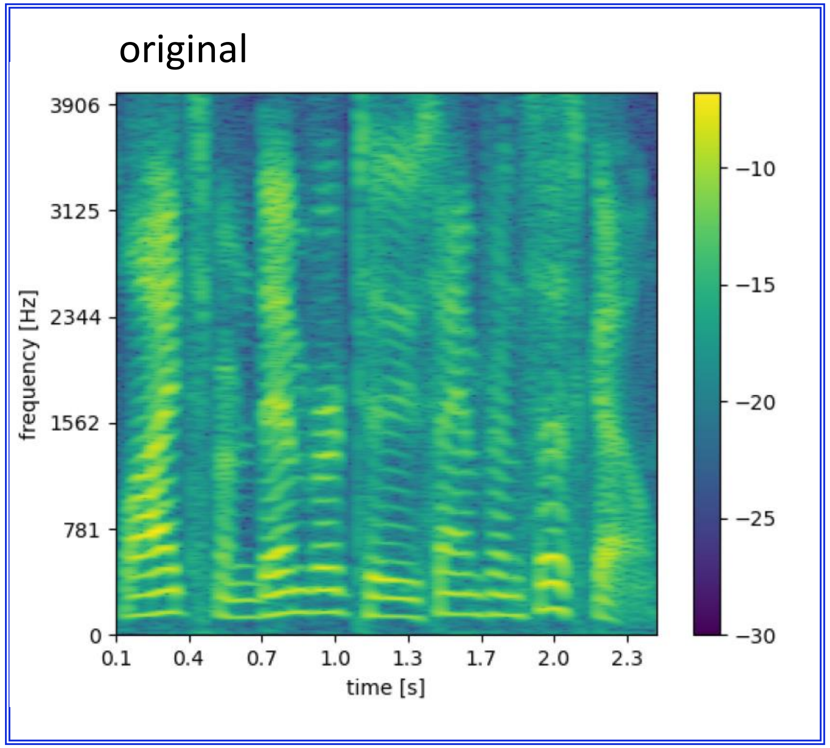
$$y[n] = x[n] + \sum_{k=1}^{P} a_k y[n-k]$$

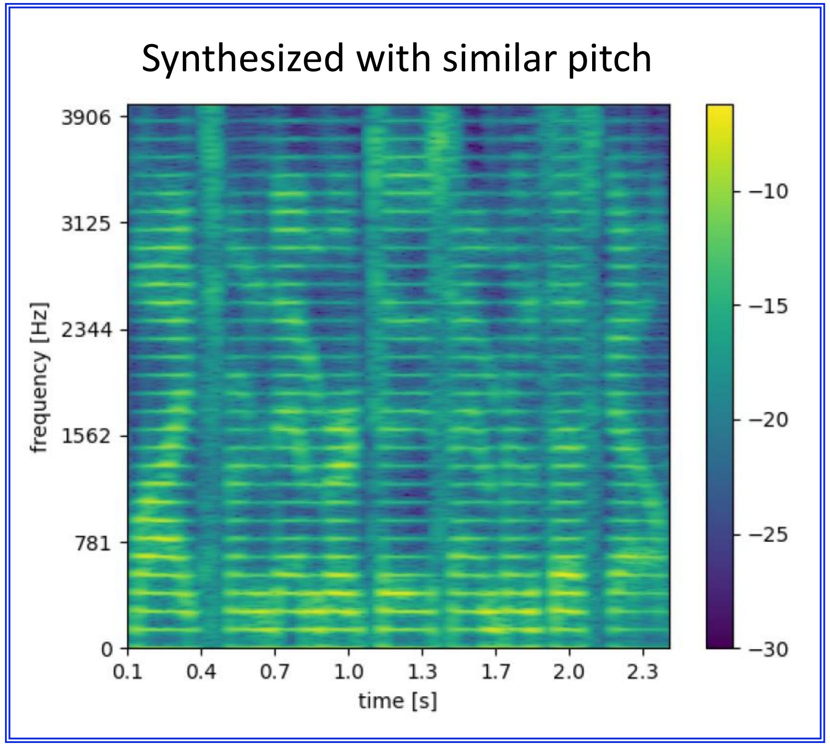
$$Y(\Omega) = X(\Omega) + \sum_{k=1}^{P} a_k e^{-j\Omega k} Y(\Omega)$$

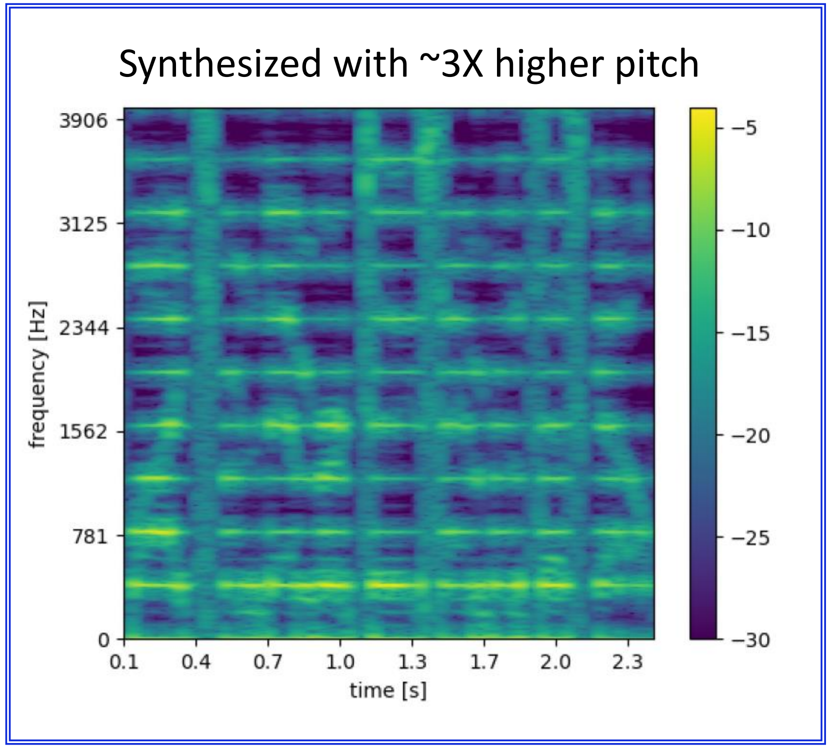
Since $Y(\Omega) = X(\Omega)H(\Omega)$

$$H(\Omega) = \frac{Y(\Omega)}{X(\Omega)} = \frac{1}{1 - \sum_{k=1}^{P} a_k e^{-j\Omega k}}$$

Synthesizing Speech Using LPC Model

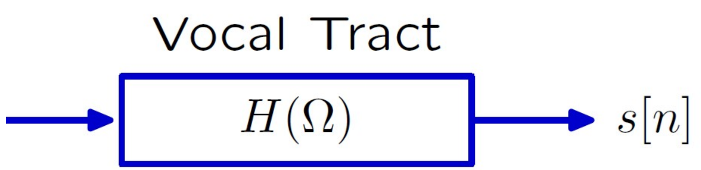




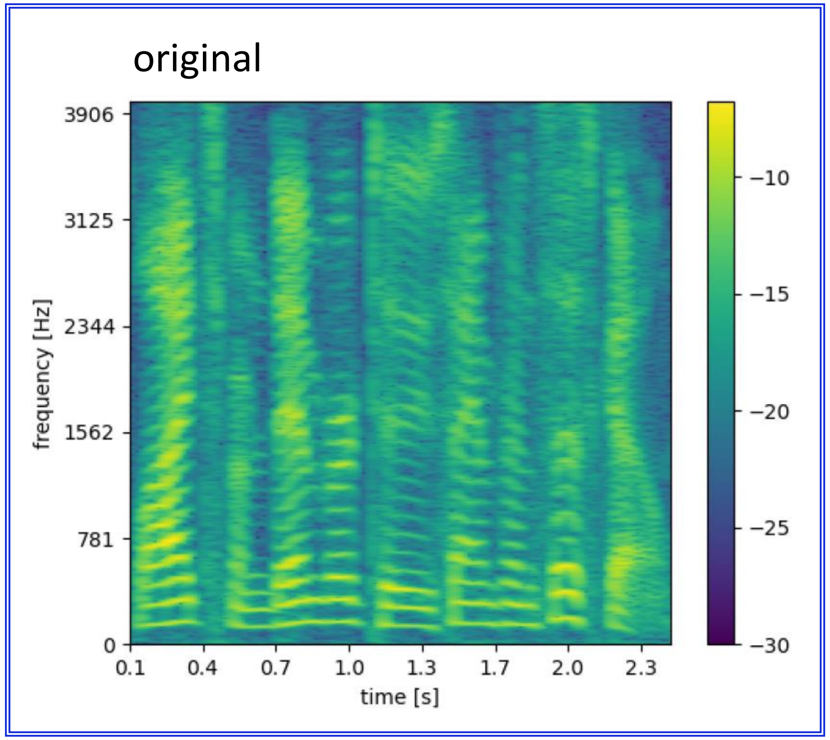


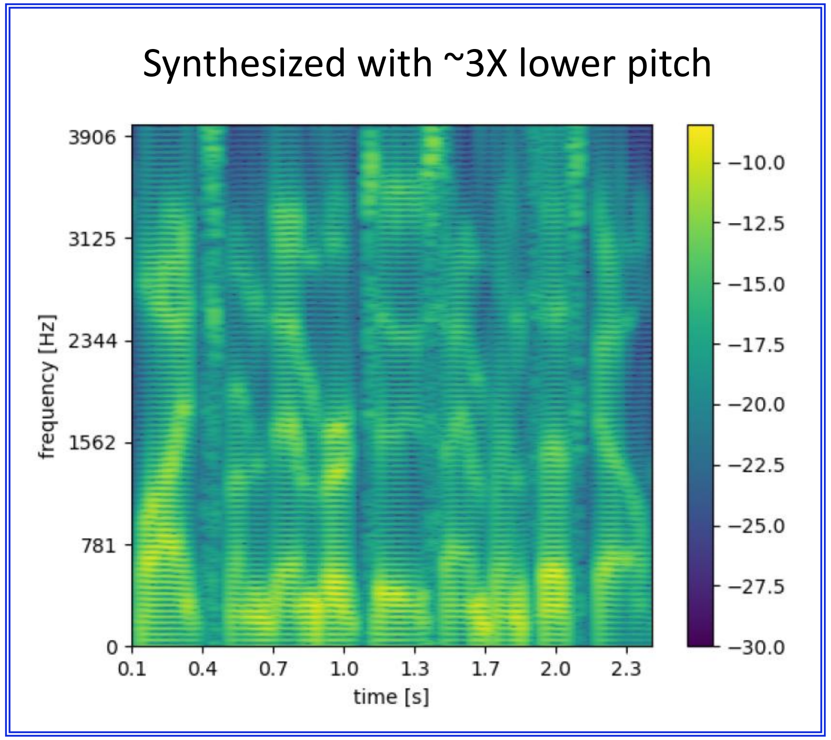
Summary of LPC procedure:

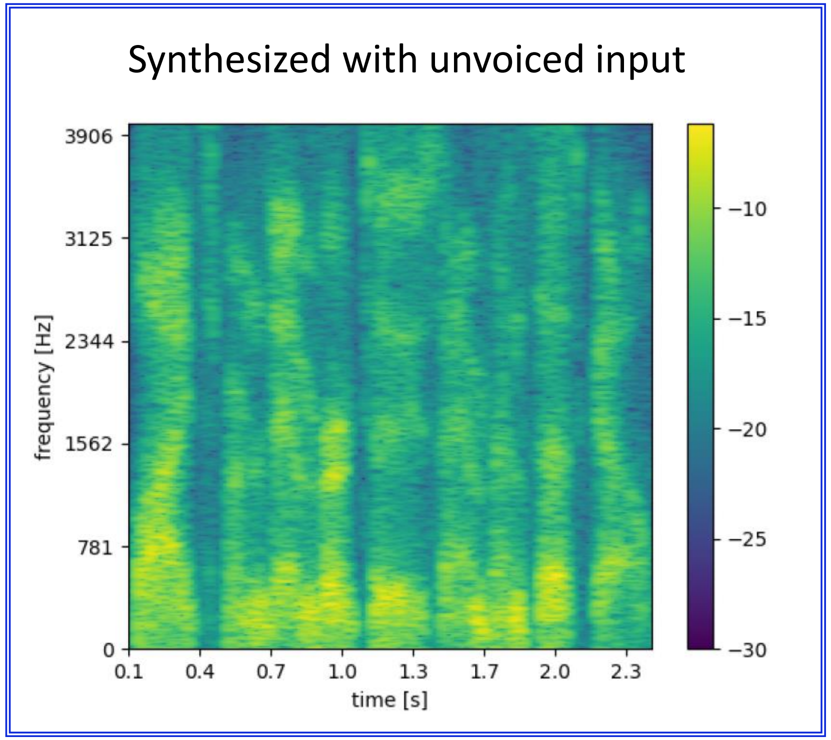
1. select a region of time using a window function w[n]
2. calculate the autocorrelation function R[i]
3. solve the set of linear equations to find a_k.

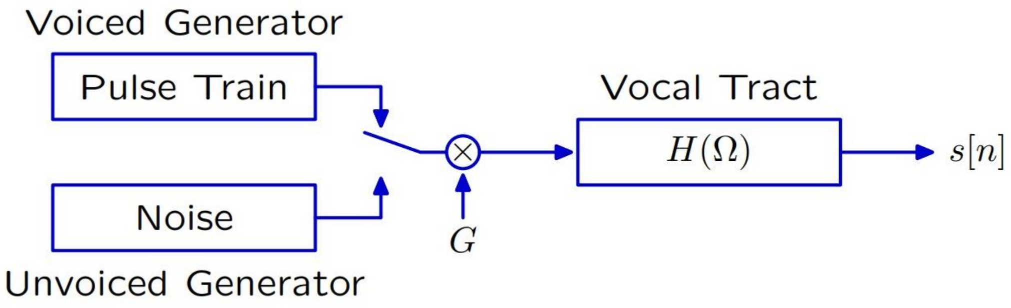


Synthesizing Speech Using LPC Model



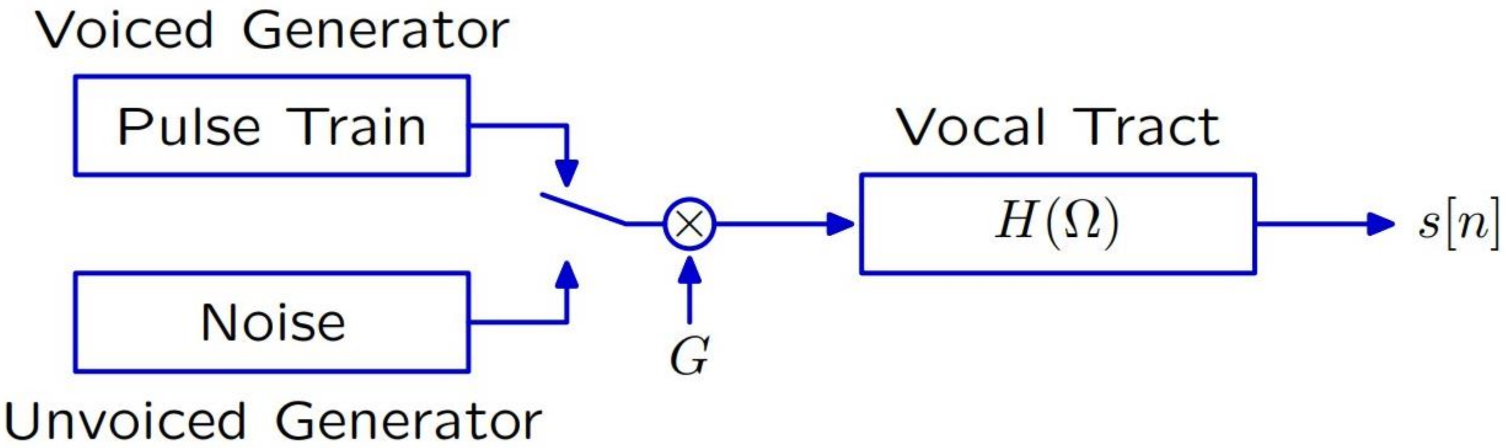






Summary

Speech production can be modeled as a source/filter model.



Frequency response of the vocal tract can be modeled with Linear predictive coding (LPC), which is widely used in in audio signal processing and speech processing.

We will now go to 4-370 for recitation & common hour