SPRING 2023

CS 378: INTRO TO SPEECH AND AUDIO PROCESSING

The Acoustic Theory of Speech Production

DAVID HARWATH
Assistant Professor, UTCS

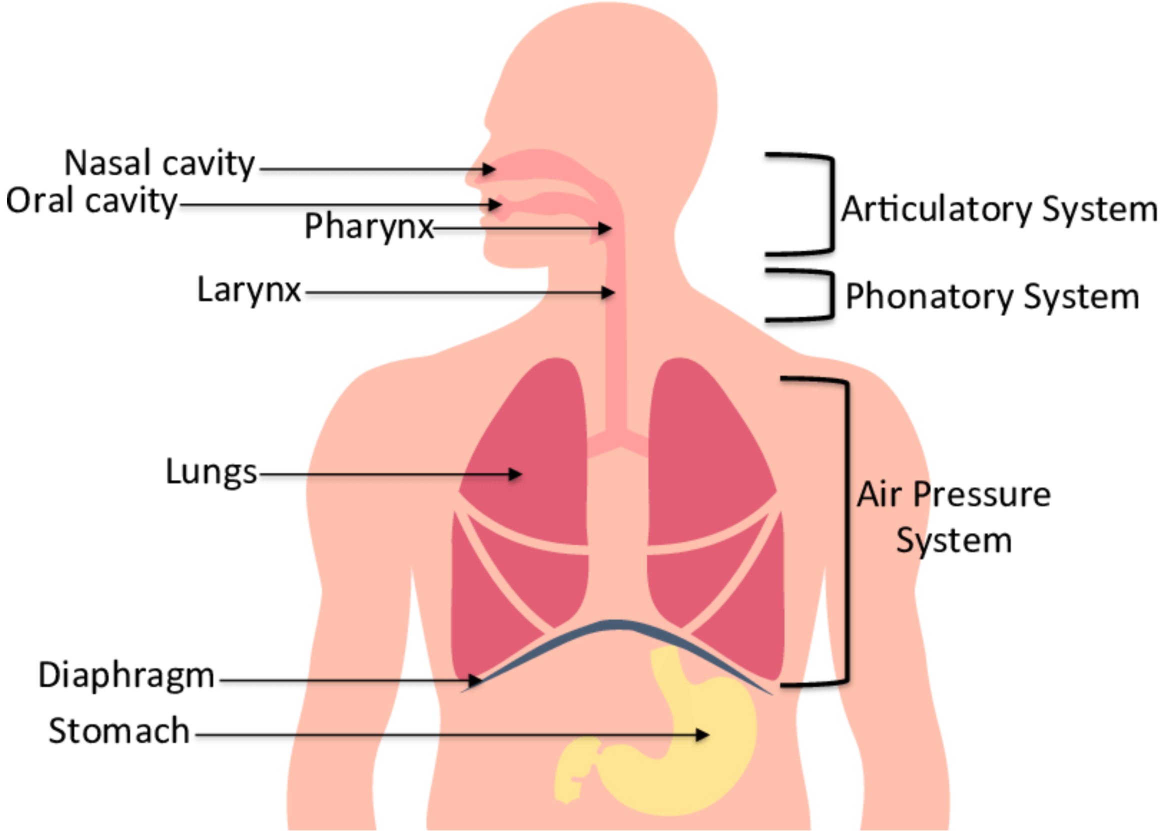
The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

Today’s agenda

- Overview of human speech production
- Acoustic tubes
- Modeling the human vocal tract with concatenated acoustic tubes

- After today, you should be able to complete exercise 1 on problem set 1

Source-Filter model of speech



Filter $H(\Omega)$

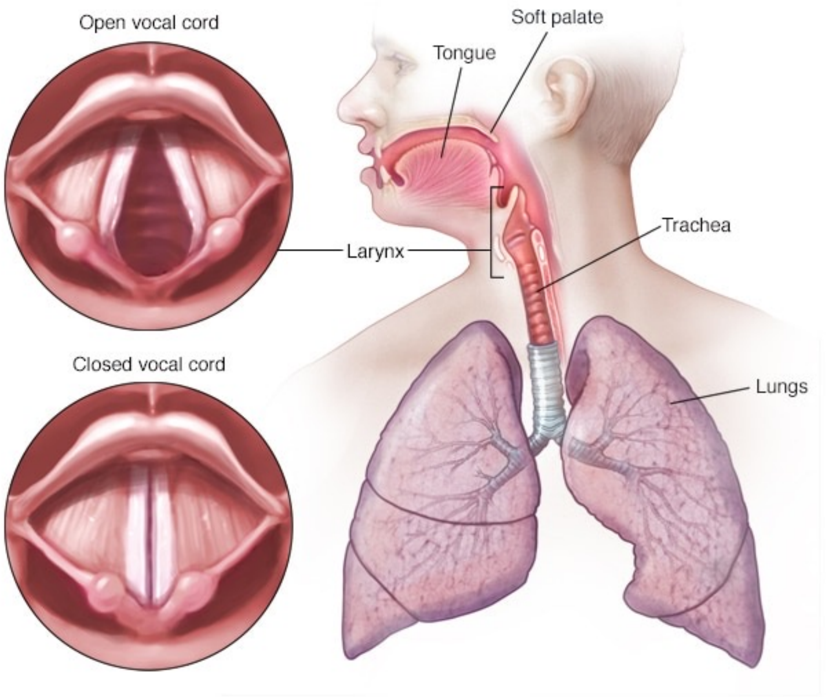
Sound source $X(\Omega)$

The sound source: your vocal cord

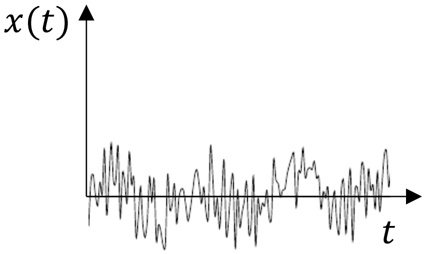
Two primary modes of operation:

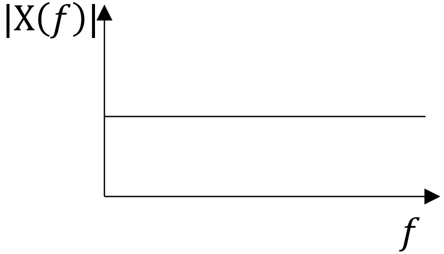
1. Unvoiced speech (open vocal cord). Produces turbulent airflow, as heard in sounds such as “ssss”, “shhh”, ‘fffff”, etc.

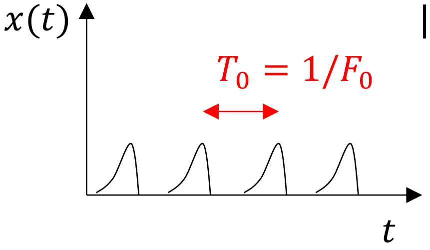
2. Voiced speech (closed vocal cord). Produces periodic (i.e. pitched) excitation, as heard in vowels

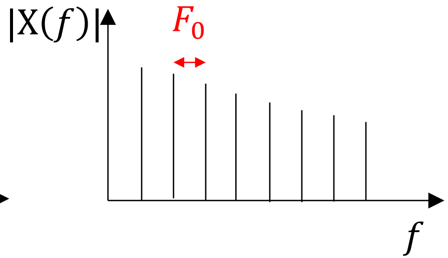


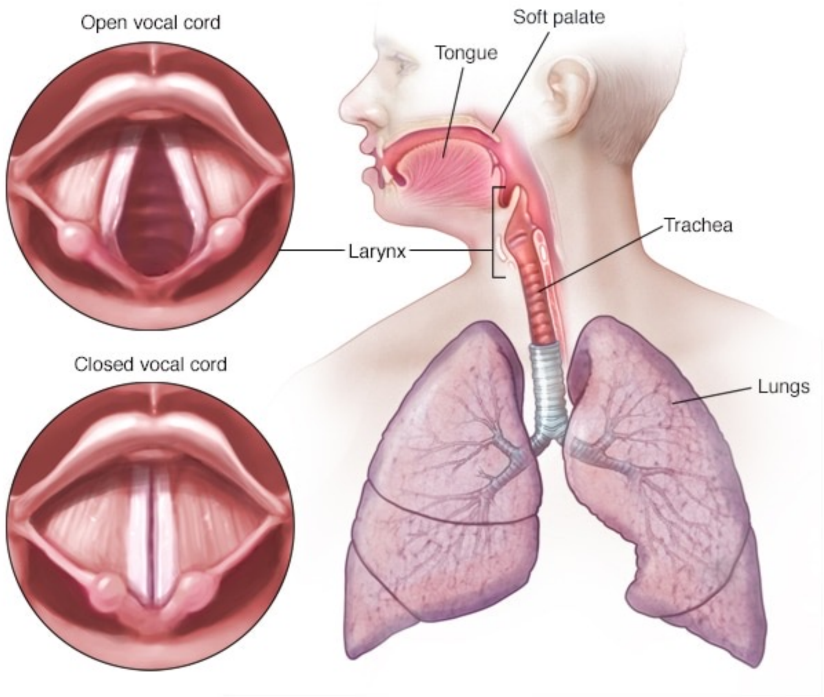
The sound source: your vocal cord







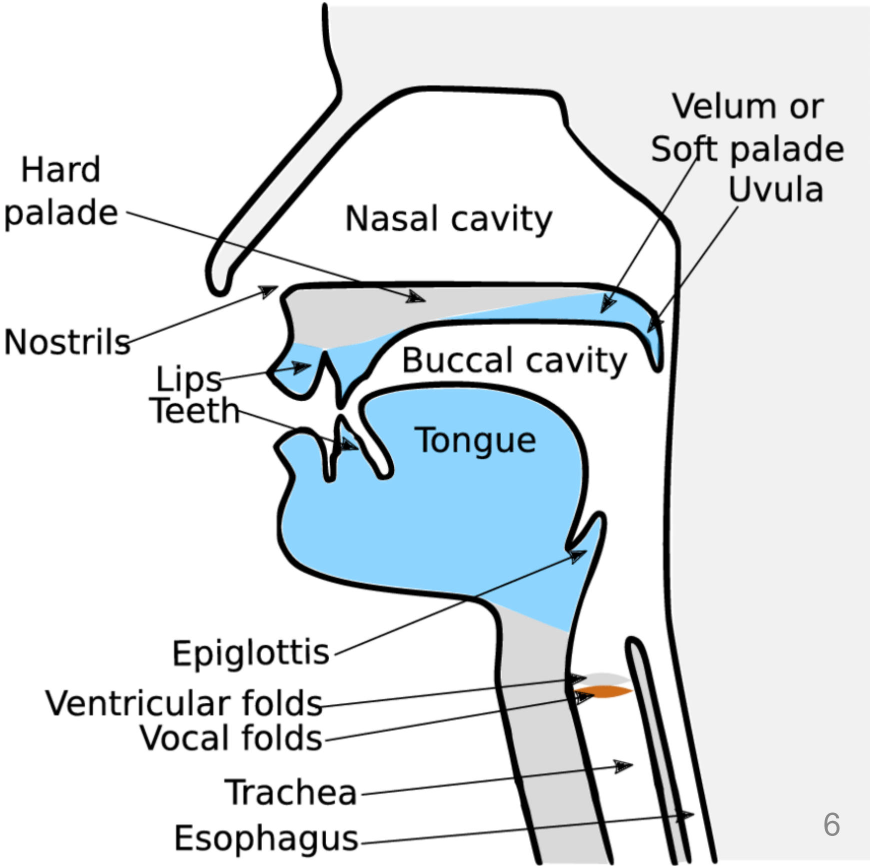




The filter: your vocal tract

- Manipulation of your articulators changes the shape of your vocal tract, which changes $H(\Omega)$

- Every person has a slightly different vocal tract, and thus a different voice
  - But the general patterns of speech sounds are universal



Analogy to brass instruments

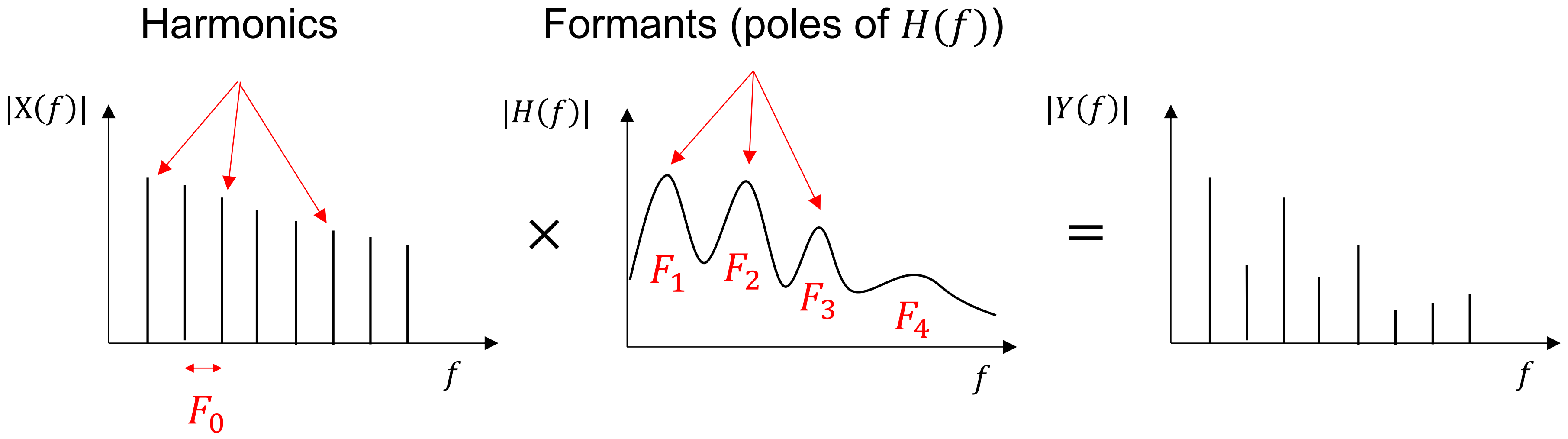
Sound Source

Embouchure (buzzing your lips)

Adjustable Filter

Changing the shape of an acoustic tube (e.g. moving the trombone's slide)

Source-Filter Speech Production

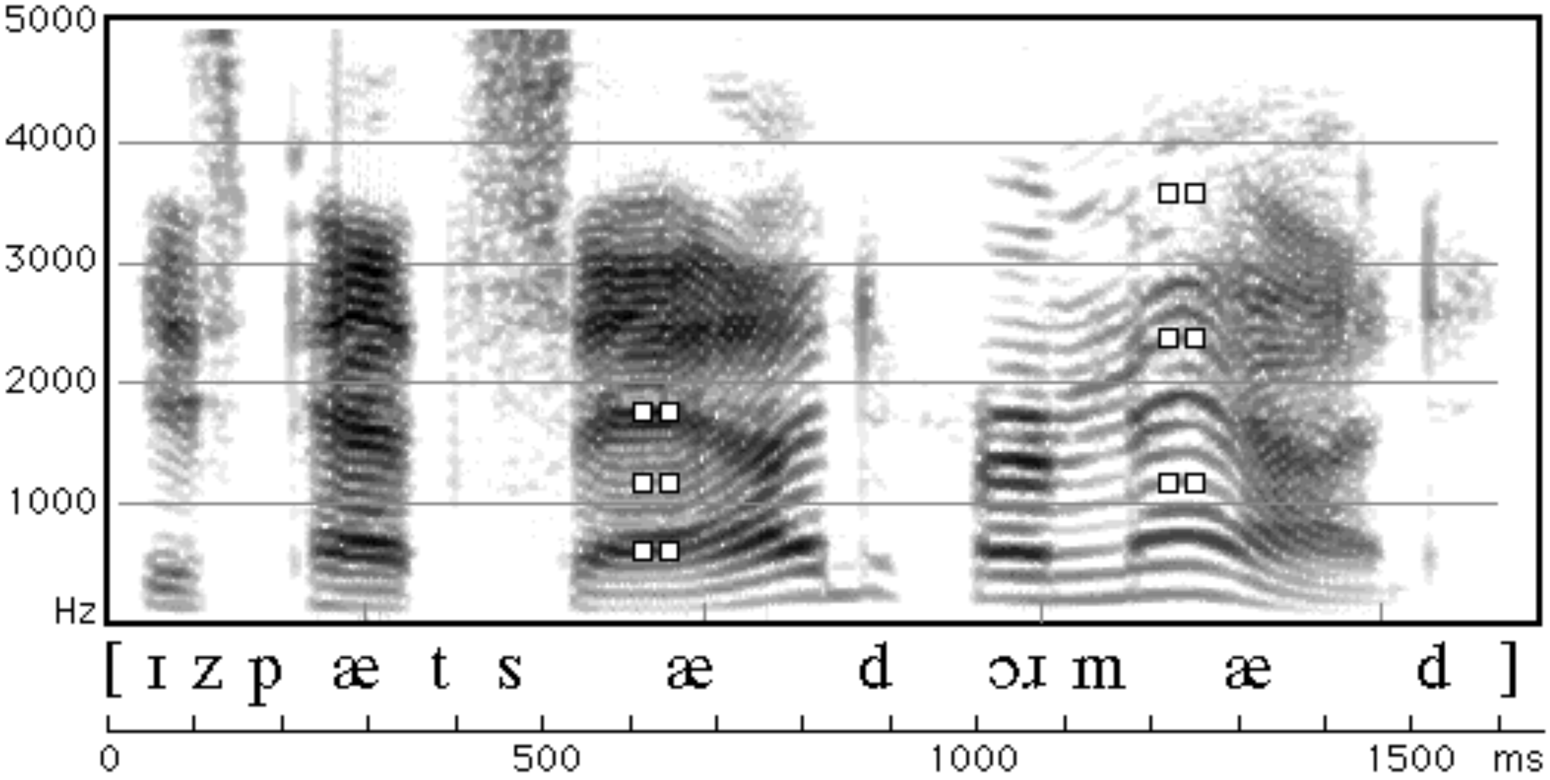


<u>Source</u>
Vocal cord vibration

<u>Filter</u>
Vocal tract frequency response

<u>Observed Spectrum</u>
What is recorded by the microphone

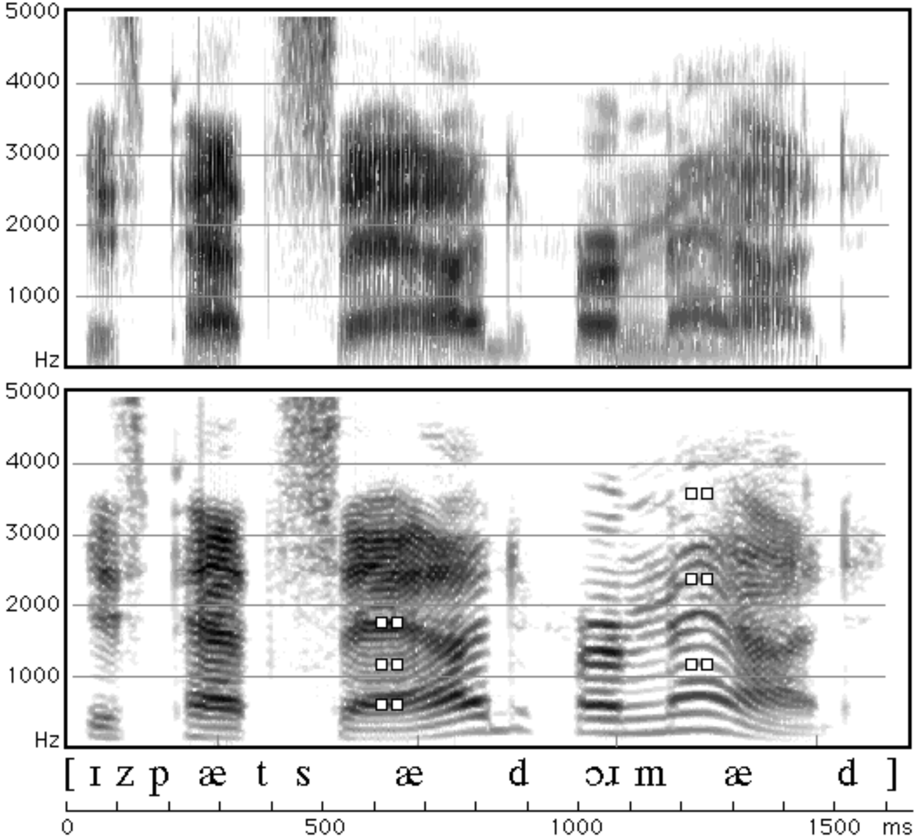
Formants and Harmonics



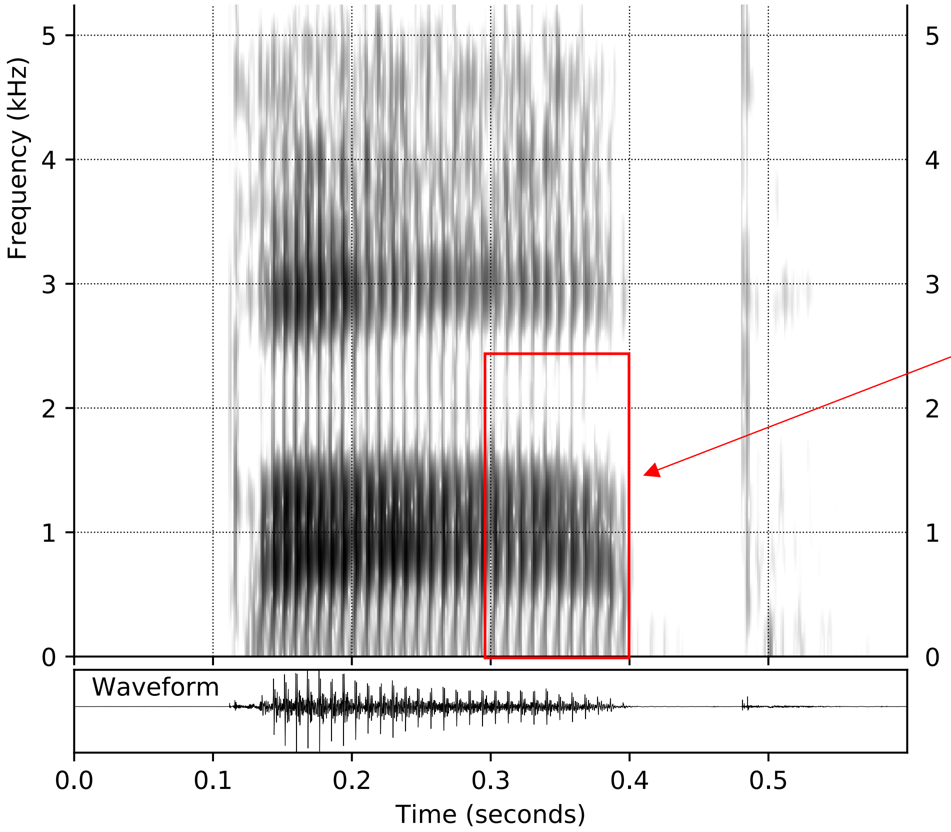
Recall: Narrowband vs. Wideband Spectrograms

Wideband spectrogram: Short STFT window blurs together harmonics, but gives sharper time detail

Narrowband spectrogram: Long STFT window reveals voicing harmonics, but with worse time detail (e.g. for stop consonants)



Estimating F0

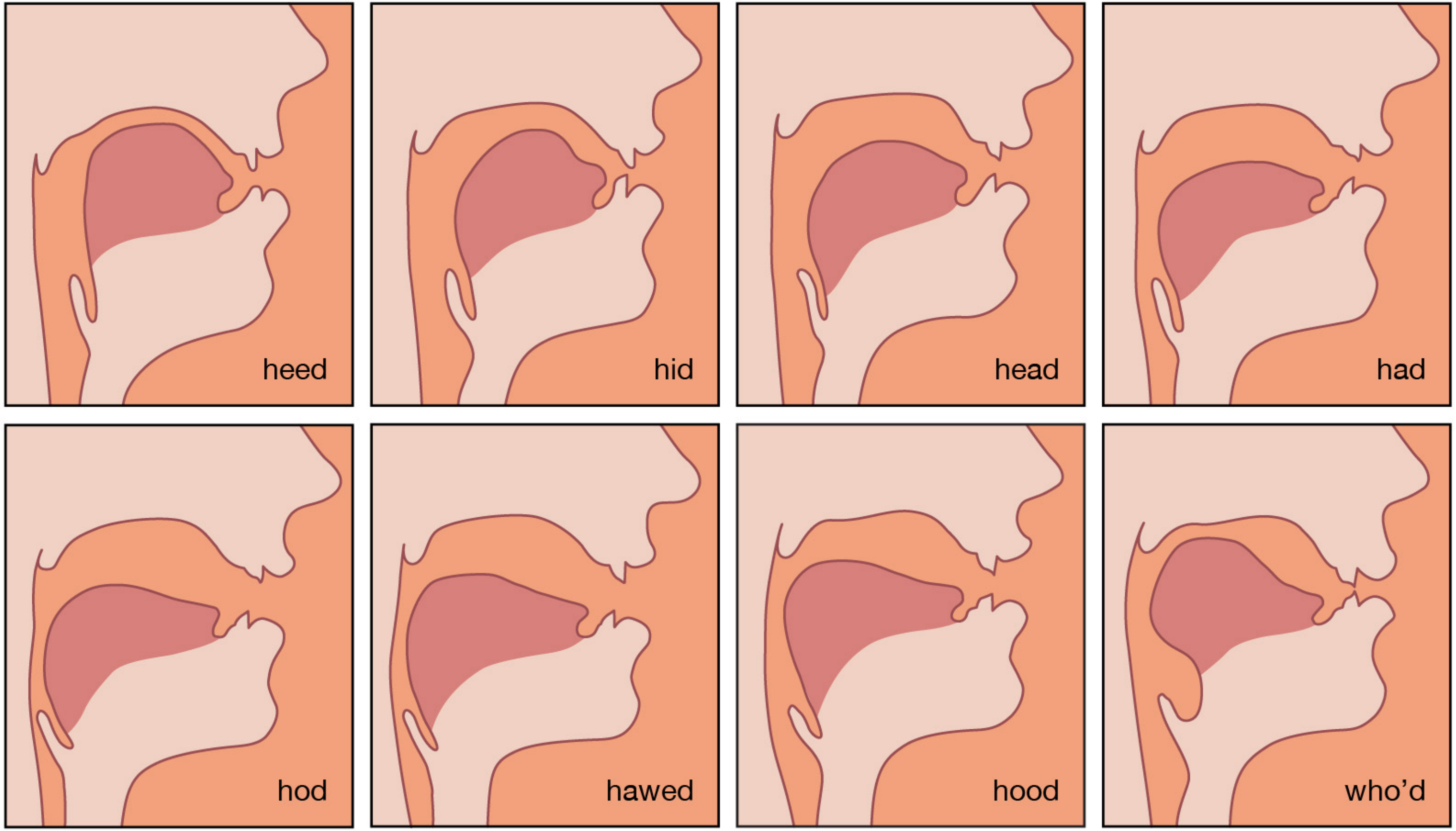


Count # pitch periods in a 0.1 (100ms) time span

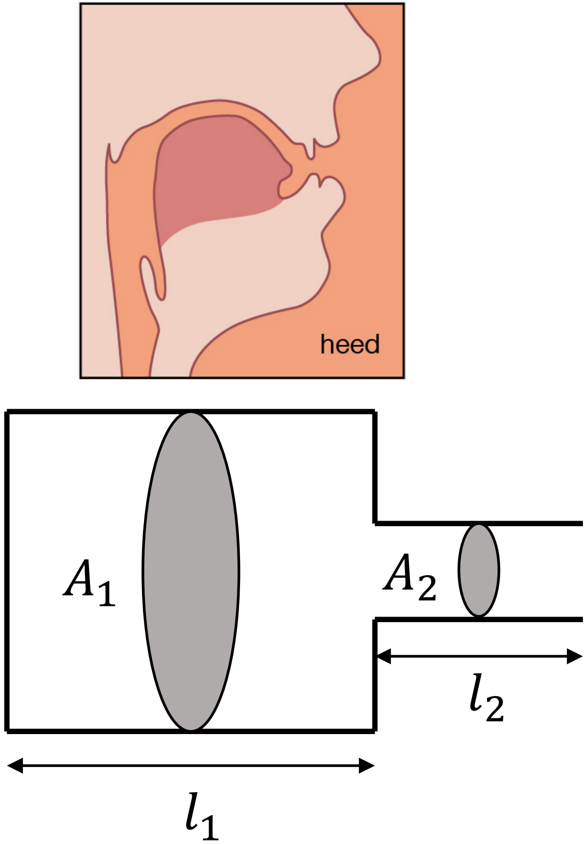
11 pitch periods / .1 seconds = 110 pitch periods per second, so F0 = 110 Hz

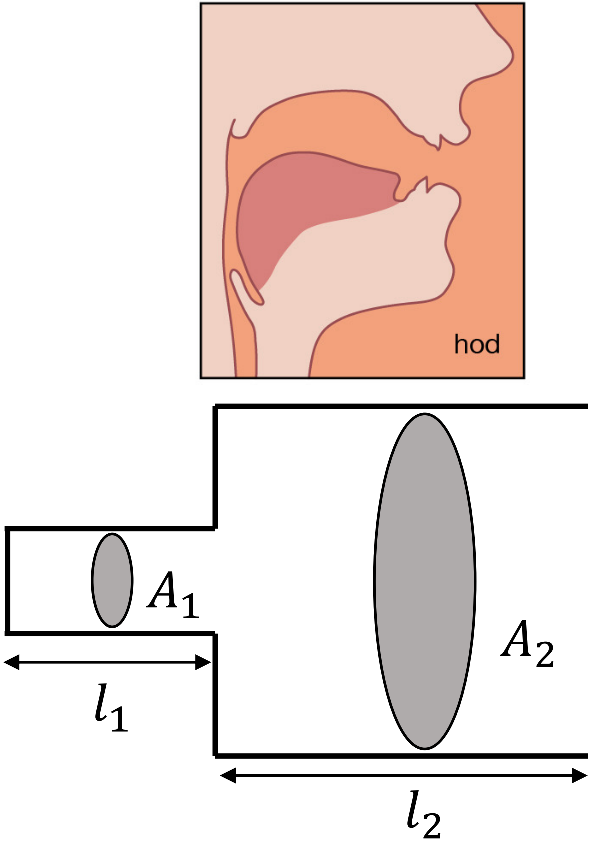
Can only do this with a wideband spectrogram (window must be shorter than a pitch period)

Vocal tract shape ⇒ vowel quality



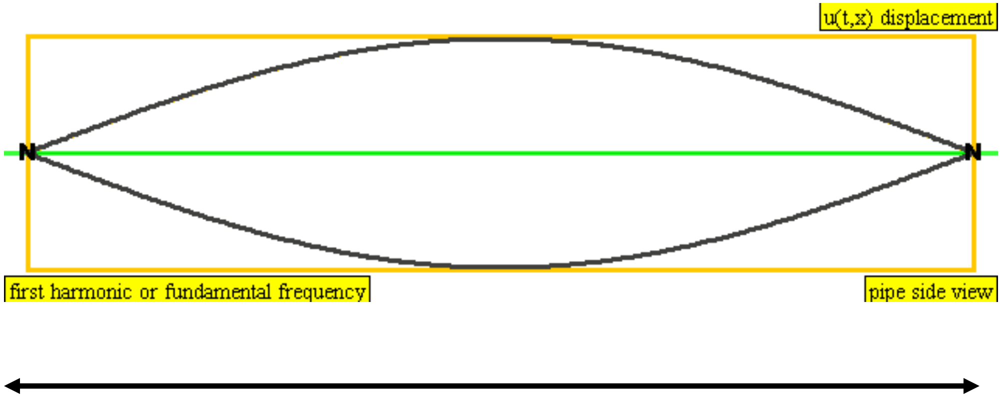
Approximating the vocal tract shape





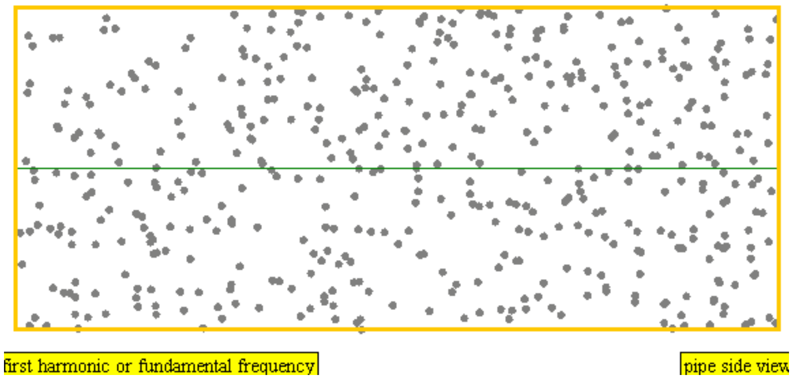
Resonances of Acoustic Tubes

Recall from physics that hollow tubes filled with air will resonate at different wavelengths depending their length



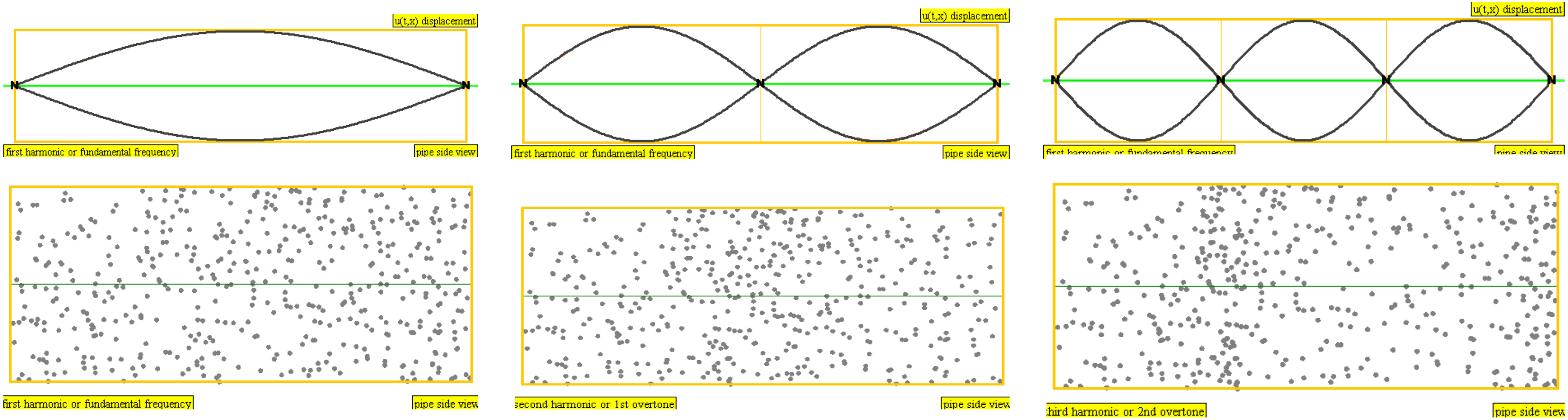
l

(Length of tube)



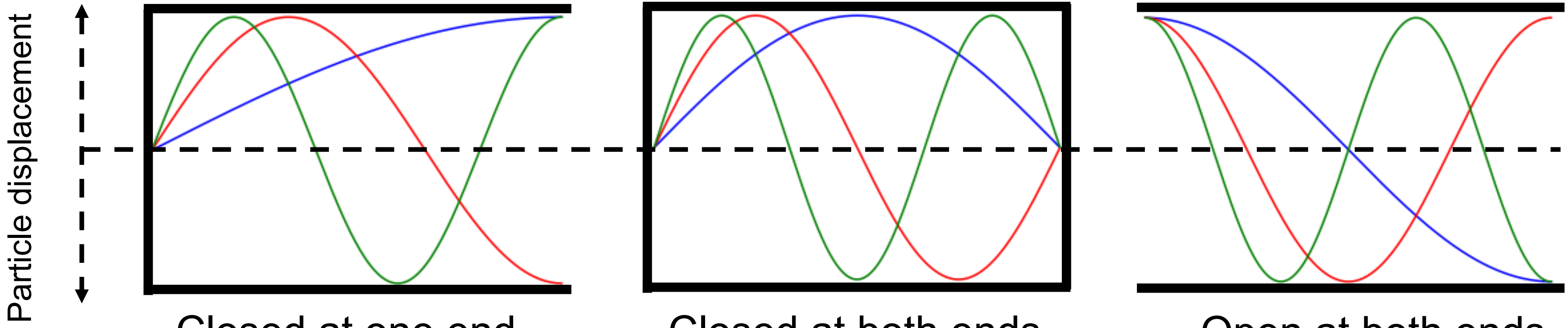
Resonances of Acoustic Tubes

An acoustic tube will always resonate at *multiple* frequencies that are harmonically related



Quarter- vs. Half-Wavelength Resonators

Resonance frequencies also depend on the boundary conditions of the tube (open vs. closed at the ends). For a resonance, particle displacement will be *zero at a solid wall*, and at a *maximum at an open end.*



Closed at one end

Can “squeeze” $\frac{1}{4}, \frac{3}{4}, \frac{5}{4}, \ldots$ of a wavelength into tube

Closed at both ends

Can “squeeze” $\frac{1}{2}, 1, \frac{3}{2}, \ldots$ of a wavelength into tube

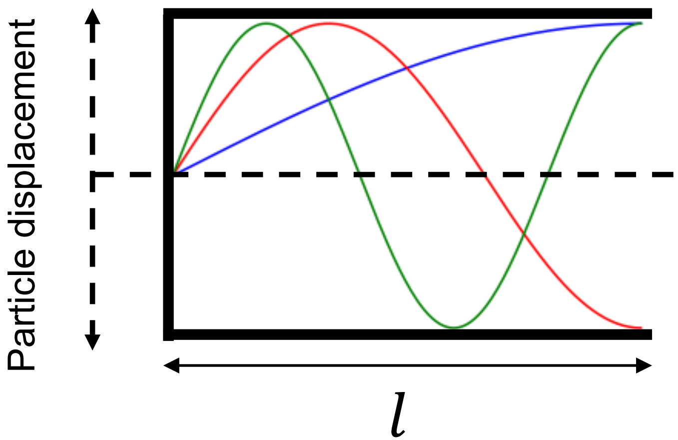
Open at both ends

Can “squeeze” $\frac{1}{2}, 1, \frac{3}{2}, \ldots$ of a wavelength into tube

“Quarter-wavelength resonator”

“Half-wavelength resonator”

Quarter-Wavelength Resonators



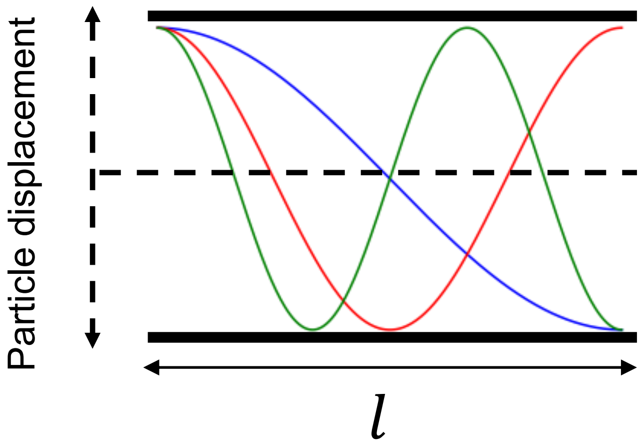
A quarter-wavelength resonator is *closed only at one end* and will have resonances at frequencies f_n given by:

$$f_n = \frac{c}{4l}(2n-1), \qquad n = 1, 2, 3, \ldots$$

Where c is the velocity of sound in air (34,000 cm/s), l is the length of the tube in cm, and f_n is in Hertz (Hz)

Note: “resonances” = “poles” = “natural frequencies”

Half-Wavelength Resonators

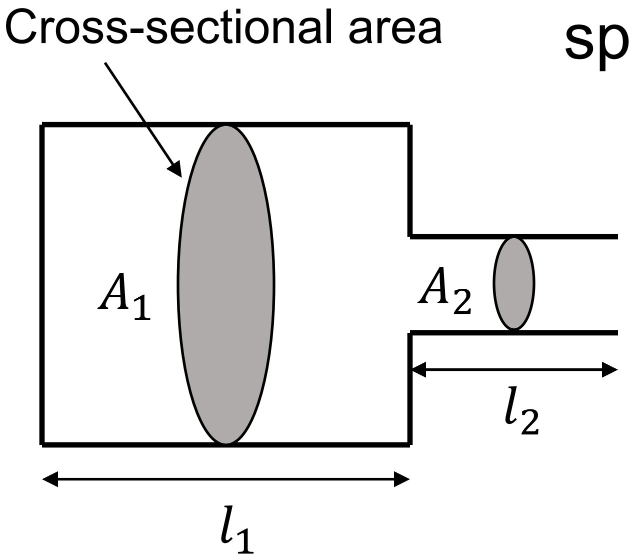


A half-wavelength resonator is *closed at both ends* ***or*** *open at both ends* and will have resonances at frequencies f_n given by:

$$f_n = \frac{c}{2l} n, \qquad n = 1, 2, 3, \ldots$$

Again where c is the velocity of sound in air (34,000 cm/s), l is the length of the tube in cm, and f_n is in Hertz (Hz)

Helmholtz Resonators



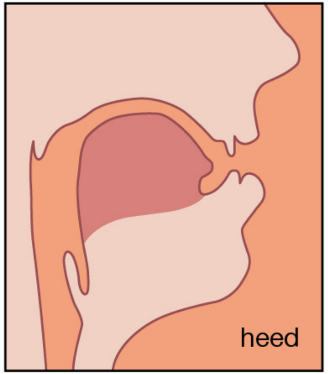
A third type of resonator that comes up in speech production is the Helmholtz resonator

It has a characteristic “bottle” shape, and has a special low frequency resonance (the *Helmholtz resonance*) at

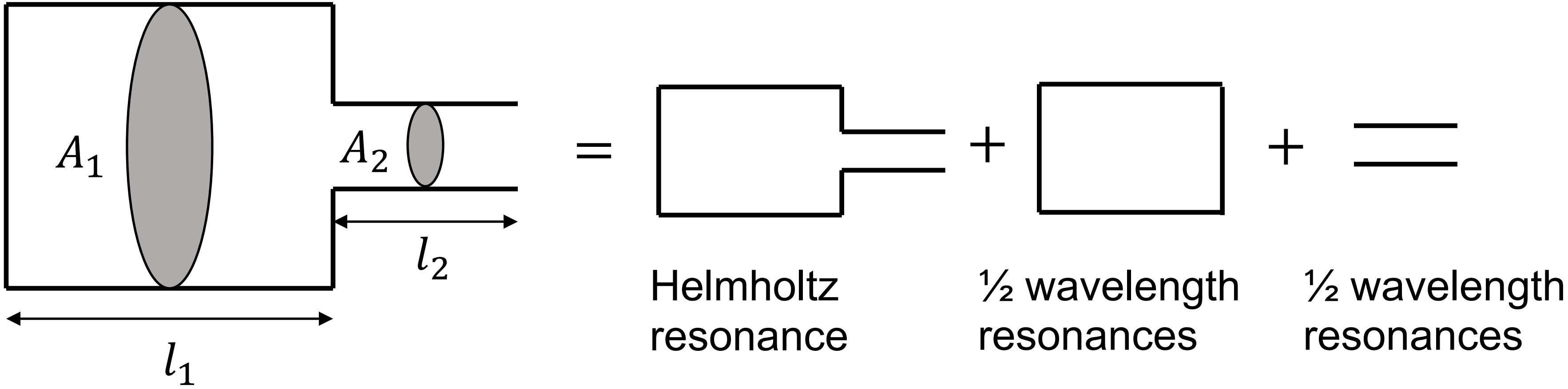
$$f = \frac{c}{2\pi}\left[\frac{A_2}{A_1 l_1 l_2}\right]^{1/2}$$

Helmholtz Demo

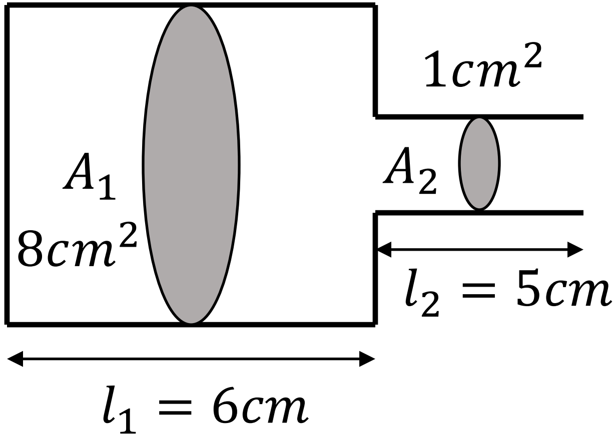
Decoupling Concatenated Tubes



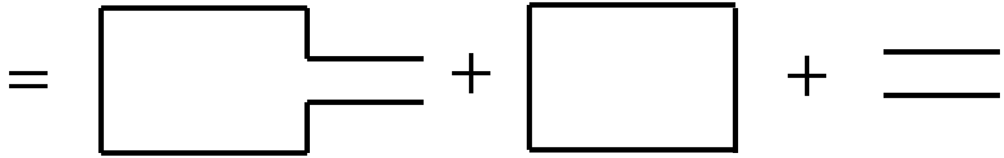
When $A_1 \gg A_2$ or $A_1 \ll A_2$, we can decouple the tubes and compute their resonances independently.
The *union* of the sets of resonances belonging to all tubes determine the *formant frequencies.*



Example: Formants for [i]

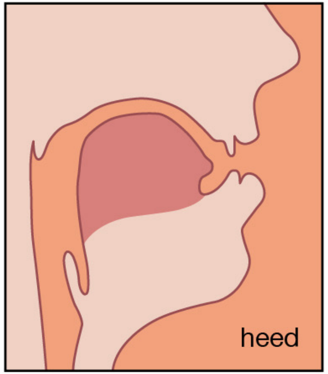


Helmholtz ½ WL ½ WL



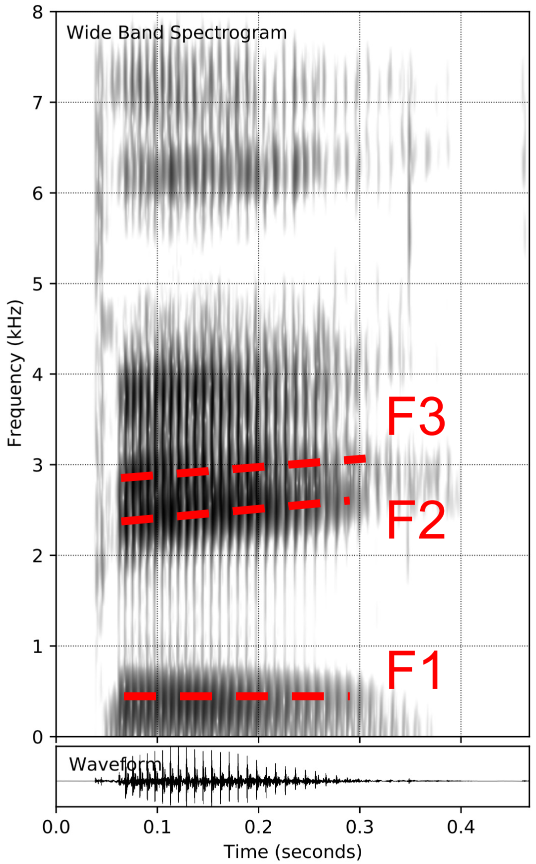
$$f = \frac{c}{2\pi}\left[\frac{A_2}{A_1 l_1 l_2}\right]^{1/2} \qquad f_n = \frac{c}{2l_1}n \qquad f_n = \frac{c}{2l_2}n$$

F1 349 Hz	F2 2833 Hz	F3 3400 Hz
	5667 Hz	6800 Hz
	8500 Hz	10200 Hz
	⋮	⋮

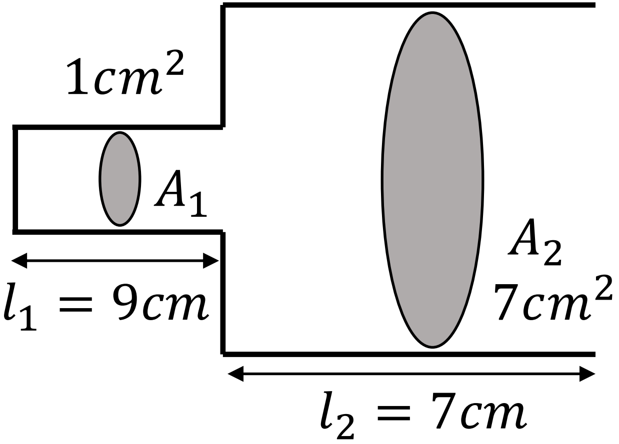


We take the union of all resonances from all tubes, sort them in ascending order, and label them as the first formant (F1), second formant (F2), and so on.

Example Spectrogram



Example: Formants for [ɑ]



=

¼ WL

$$f_n = \frac{c}{4l}(2n-1)$$

F1 944 Hz

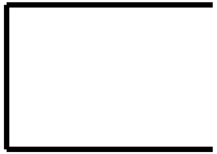
F3 2833 Hz

4722 Hz

⋮

+

¼ WL



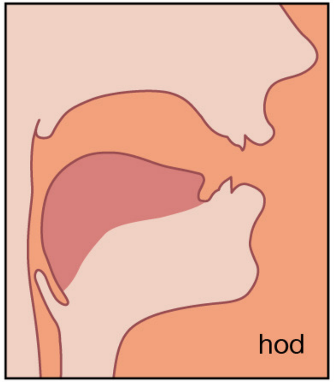
$$f_n = \frac{c}{4l}(2n-1)$$

F2 1214 Hz

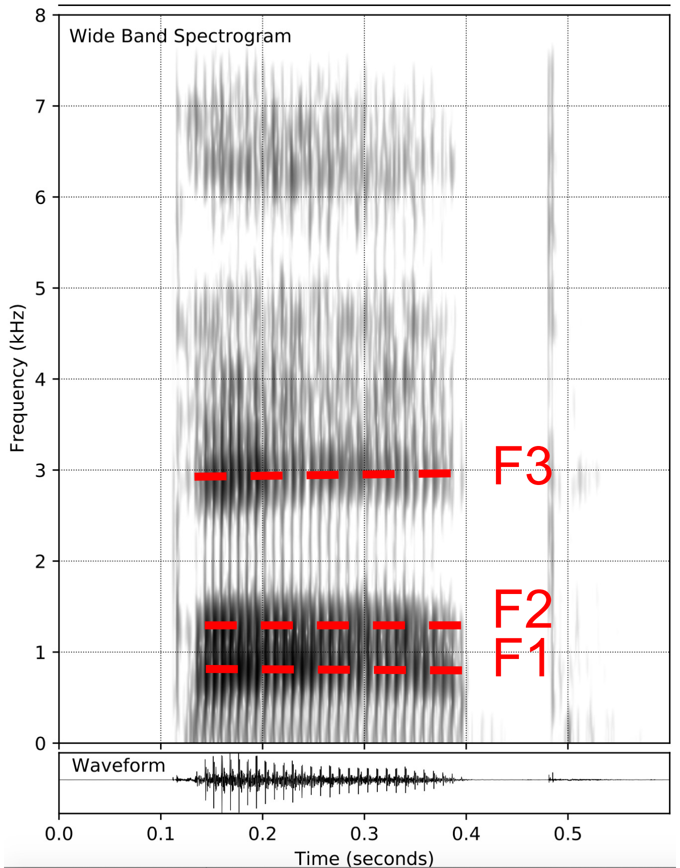
3642 Hz

6071 Hz

⋮

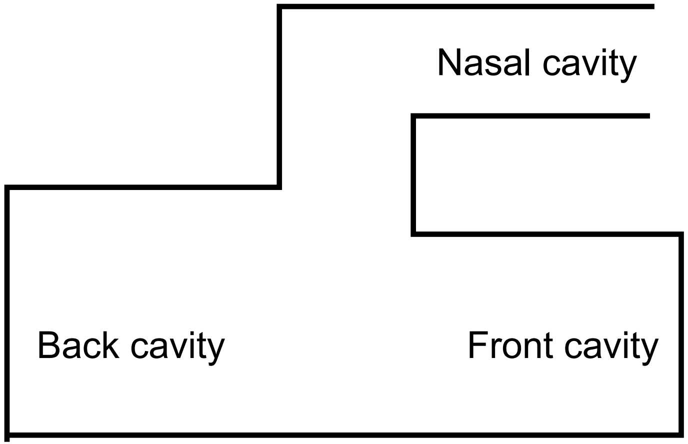


Example Spectrogram



Opening the nasal cavity

When you make a nasal consonant like an ”m” or “n”, you lower your velum which couples your nasal cavity to your vocal tract



Resonances of the front cavity “trap” acoustic energy and prevent it from radiating out from the nasal cavity

This gives rise to zeros in the transfer function from the vocal folds to the nostrils, which cancel out formants

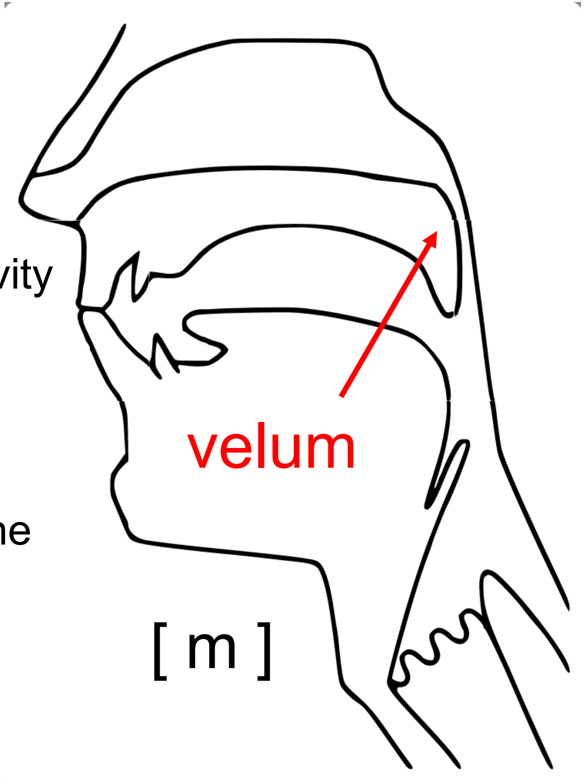
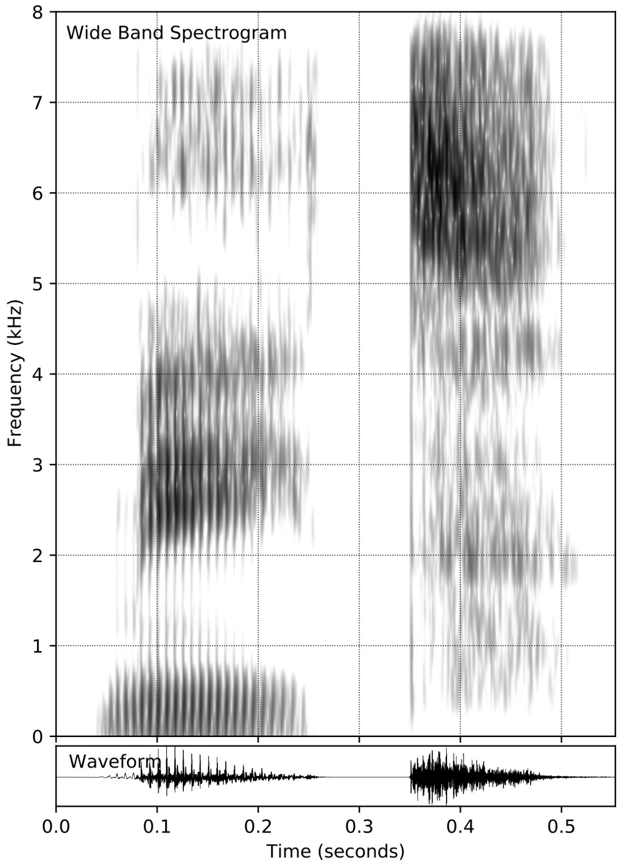
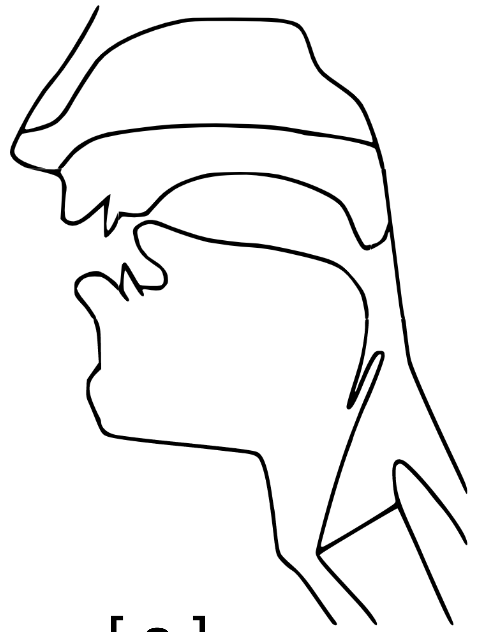


Illustration created with http://smu-facweb.smu.ca/~s0949176/sammy/

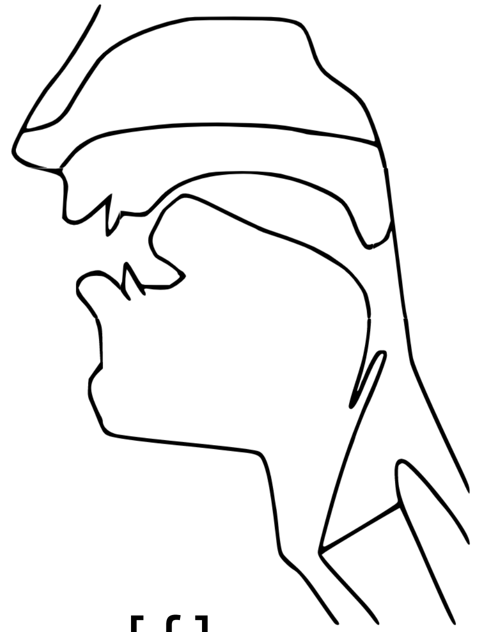
Example Spectrogram



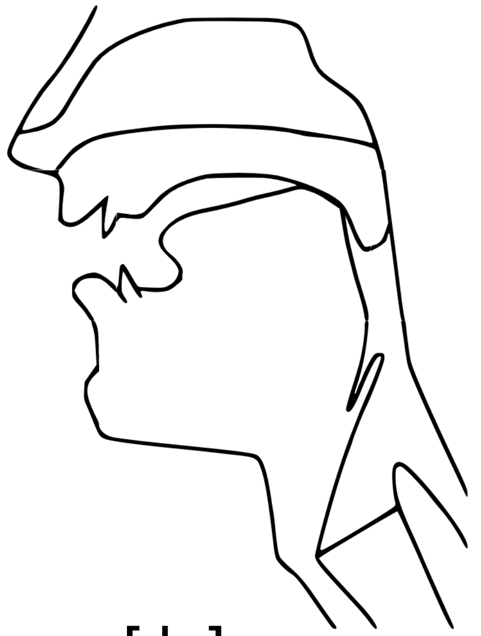
Constricting for Consonants



[s]

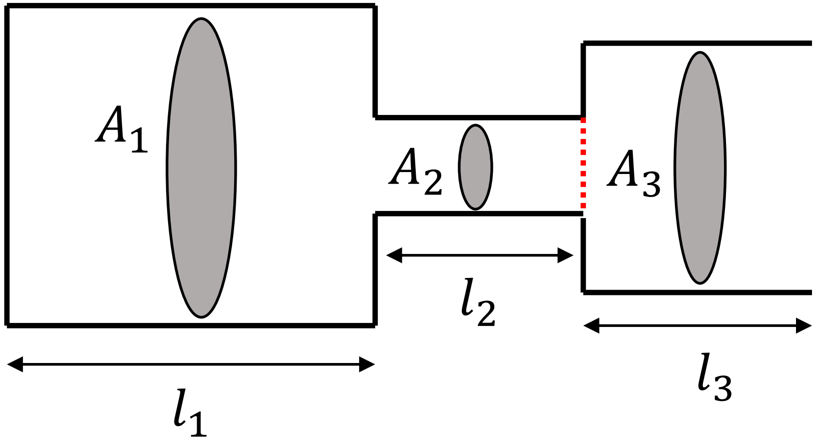


[ʃ]

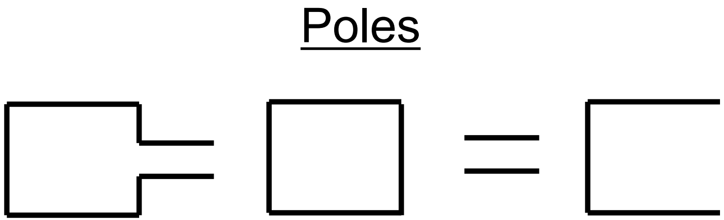


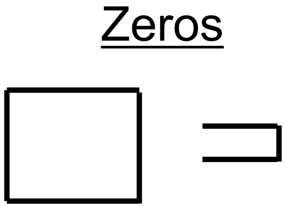
[k]

Constricting for Consonants

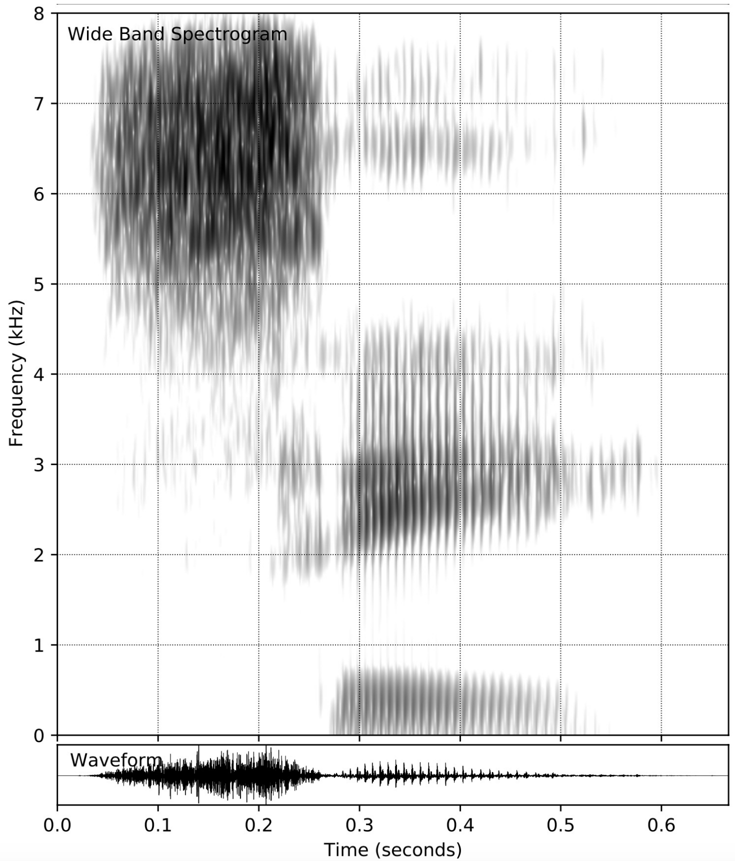


A constriction will also introduce zeros that correspond to the resonances of the tubes behind the very front of the constriction, where we also treat the front of the constriction as a *hard wall*





Example Spectrogram



Rounding the lips

Rounding your lips has the effect of slightly *increasing* the length of the vocal tract, and drags all formants down