CS 224S / LINGUIST 281
Speech Recognition, Synthesis, and Dialogue

Dan Jurafsky

Lecture 1: Short introduction to the course, the ARPAbet, and Articulatory Phonetics

Today, Jan 6, Week 1

- Overview and very brief history
- Administration
  - Overview of course topics
  - Grading
- Articulatory Phonetics
- ARPAbet transcription

Applications of Speech Recognition/ Understanding (ASR/ASU)

- Dictation
- Telephone-based Information
  - Google voice search
  - Directions
  - Air travel, banking, etc
- Hands-free (in car)
- Second language ('L2') (accent reduction)
- Audio archive searching and aligning

Applications of Speech Synthesis/ Text-to-Speech (TTS)

- Games
- Telephone-based Information (directions, air travel, banking, etc)
- Eyes-free (in car)
- Reading/speaking for disabled
- Education (Reading tutors, L2)

Applications of Speaker/Lg Recognition

- Language recognition for call routing
- Speaker Recognition:
  - Speaker verification (binary decision)
    - Voice password, telephone assistant
  - Speaker identification (one of N)
    - Criminal investigation

One example: Extraction of Social Meaning from Speech

- Detection of student uncertainty in tutoring
  - Forbes-Riley et al. (2008)
- Emotion detection (annoyance)
  - Ang et al. (2002)
- Detection of deception
  - Newman et al. (2003)
- Detection of charisma
  - Rosenberg and Hirschberg (2005)
- Speaker stress, trauma
  - Rude et al. (2004), Pennebaker and Lay (2002)

Conversational style

- Given speech and text from a conversation
- Can we tell if a speaker is
  - Awkward?
  - Flirtatious?
  - Friendly?
- Dataset:
  - 1000 4-minute "speed-dates"
  - Each subject rated their partner for these styles
  - The following segment has been lightly signal-processed:

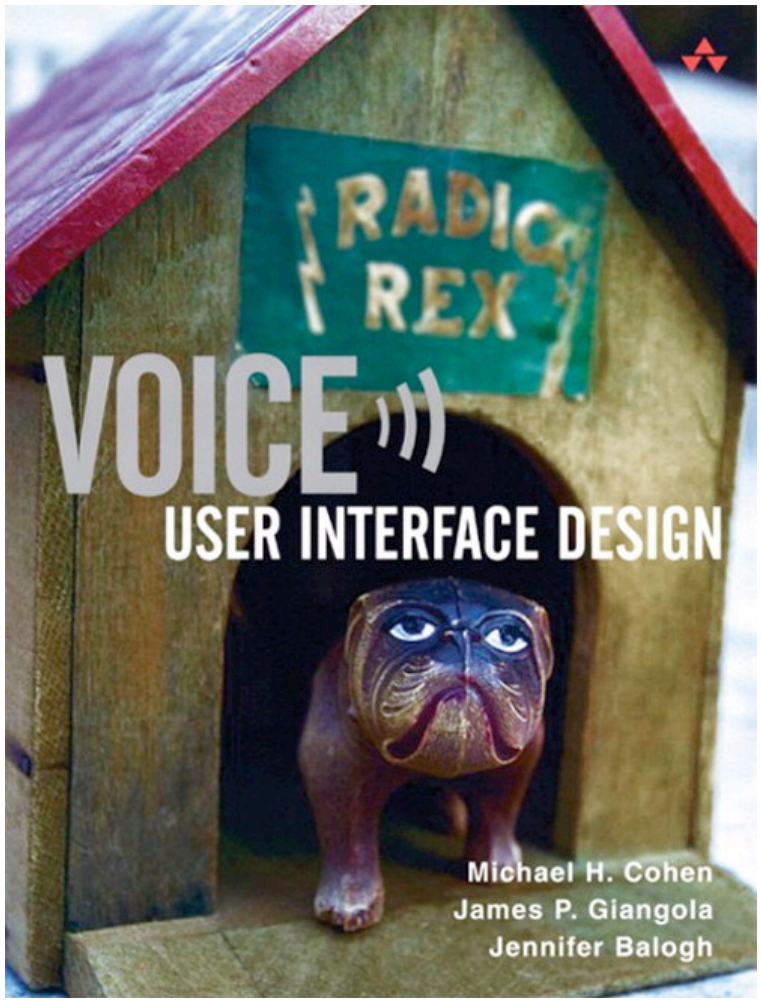
History: foundational insights 1900s-1950s

- Automaton:
  - Markov 1911
  - Turing 1936
  - McCulloch-Pitts neuron (1943)
    - http://marr.bsee.swin.edu.au/~dtl/het704/lecture10/ann/node1.html
    - http://diwww.epfl.ch/mantra/tutorial/english/mcpits/html/
  - Shannon (1948) link between automata and Markov models
- Human speech processing
  - Fletcher at Bell Labs (1920's)
- Probabilistic/Information-theoretic models
  - Shannon (1948)

Synthesis precursors

- Von Kempelen mechanical (bellows, reeds) speech production simulacrum
- 1929 Channel vocoder (Dudley)

History: Early Recognition



- 1920's Radio Rex
  - Celluloid dog with iron base held within house by electromagnet against force of spring
  - Current to magnet flowed through bridge which was sensitive to energy at 500 Hz
  - 500 Hz energy caused bridge to vibrate, interrupting current, making dog spring forward
  - The sound "e" (ARPAbet [eh]) in Rex has 500 Hz component

History: early ASR systems

- 1950's: Early Speech recognizers
  - 1952: Bell Labs single-speaker digit recognizer
    - Measured energy from two bands (formants)
    - Built with analog electrical components
    - 2% error rate for single speaker, isolated digits
  - 1958: Dudley built classifier that used continuous spectrum rather than just formants
  - 1959: Denes ASR combining grammar and acoustic probability
- 1960's
  - FFT - Fast Fourier transform (Cooley and Tukey 1965)
  - LPC - linear prediction (1968)
  - 1969 John Pierce letter "Whither Speech Recognition?"
    - Random tuning of parameters,
    - Lack of scientific rigor, no evaluation metrics
    - Need to rely on higher level knowledge

ASR: 1970's and 1980's

- Hidden Markov Model 1972
  - Independent application of Baker (CMU) and Jelinek/Bahl/Mercer lab (IBM) following work of Baum and colleagues at IDA
- ARPA project 1971-1976
  - 5-year speech understanding project: 1000 word vocab, continous speech, multi-speaker
  - SDC, CMU, BBN
  - Only 1 CMU system achieved goal
- 1980's+
  - Annual ARPA "Bakeoffs"
  - Large corpus collection
    - TIMIT
    - Resource Management
    - Wall Street Journal

State of the Art

- ASR
  - speaker-independent, continuous, no noise, world's best research systems:
    - Human-human speech: ~10-20% Word Error Rate (WER)
    - Human-machine speech: ~3-5% WER
- TTS (demo next week)

LVCSR Overview

- Large Vocabulary Continuous (Speaker-Independent) Speech Recognition
  - Build a statistical model of the speech-to-words process
  - Collect lots of speech and transcribe all the words
  - Train the model on the labeled speech
  - Paradigm: Supervised Machine Learning + Search

Unit Selection TTS Overview

- Collect lots of speech (5-50 hours) from one speaker, transcribe very carefully, all the syllables and phones and whatnot
- To synthesize a sentence, patch together syllables and phones from the training data.
- Paradigm: search

Requirements and Grading

- Readings:
  - Required Text:
  - Selected chapters from
    - Jurafsky & Martin, 2008. Speech and Language Processing.
    - Taylor, Paul. 2009. Text-to-Speech Synthesis.
  - Later in the course: a few conference and journal papers
- Grading
  - Homework: 45%
    - 7 assignments
  - Final Project: 45%
    - Group projects (3 people) are fine
  - Participation: 10%

Overview of the course

- http://www.stanford.edu/class/cs224s

Phonetics

- ARPAbet
  - An alphabet for transcribing American English phonetic sounds.
- Articulatory Phonetics
  - How speech sounds are made by articulators (moving organs) in mouth.
- Acoustic Phonetics
  - Acoustic properties of speech sounds

ARPAbet Vowels

	b_d	ARPA		b_d	ARPA
1	bead	iy	9	bode	ow
2	bid	ih	10	booed	uw
3	bayed	ey	11	bud	ah
4	bed	eh	12	bird	er
5	bad	ae	13	bide	ay
6	bod(y)	aa	14	bowed	aw
7	bawd	ao	15	Boyd	oy
8	Budd(hist)	uh			

Note: Many speakers pronounce Buddhist with the vowel uw as in booed, So for them [uh] is instead the vowel in "put" or "book"

ARPAbet

- http://www.stanford.edu/class/cs224s/arpabet.html

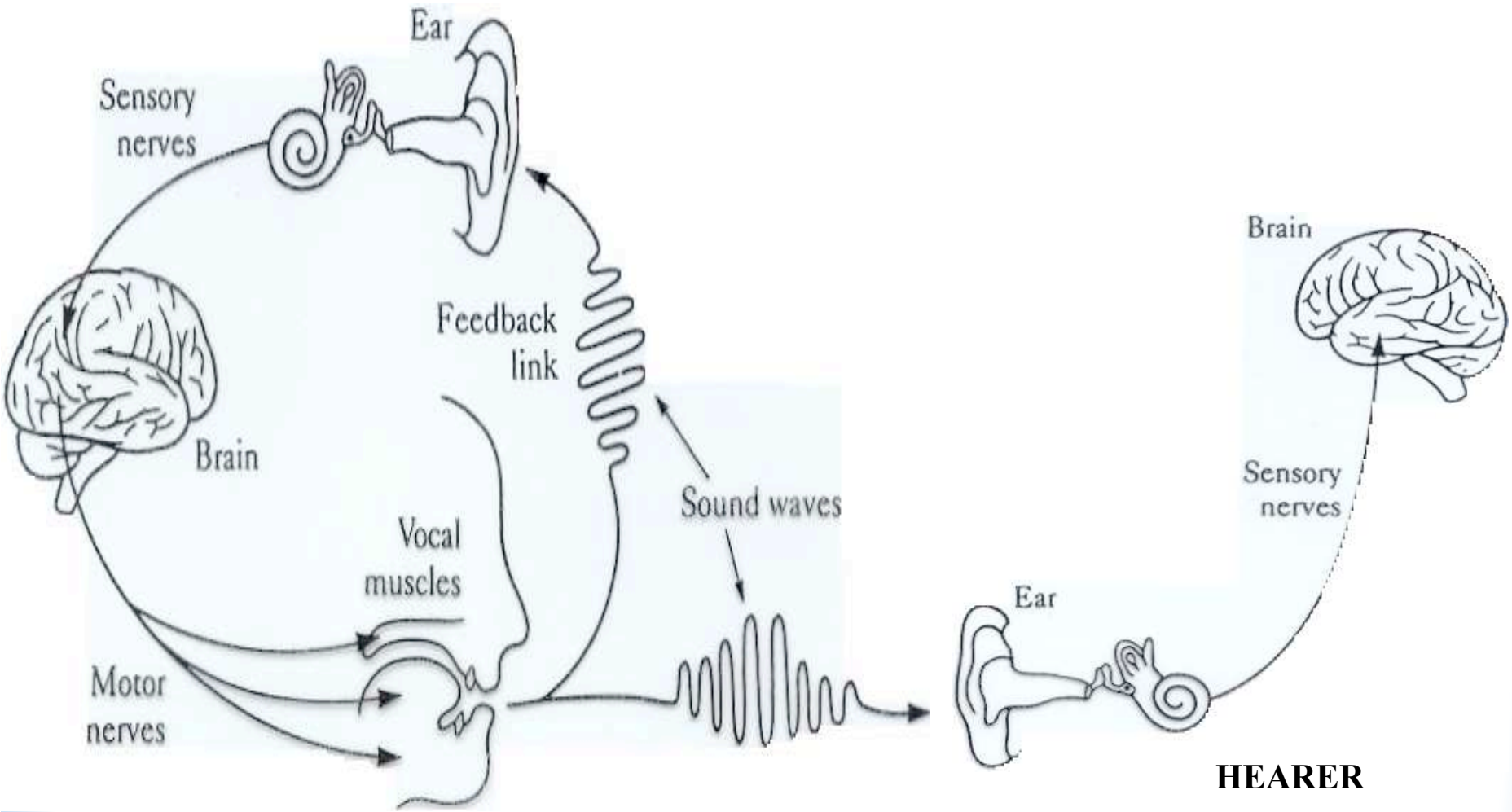
ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[p]	[p]	parsley	[p aa r s l iy]
[t]	[t]	tea	[t iy]
[k]	[k]	cook	[k uh k]
[b]	[b]	bay	[b ey]
[d]	[d]	dill	[d ih l]
[g]	[g]	garlic	[g aa r l ix k]
[m]	[m]	mint	[m ih n t]
[n]	[n]	nutmeg	[n ah t m eh g]
[ng]	[ŋ]	baking	[b ey k ix ng]
[f]	[f]	flour	[f l aw axr]
[v]	[v]	clove	[k l ow v]
[th]	[θ]	thick	[th ih k]
[dh]	[ð]	those	[dh ow z]
[s]	[s]	soup	[s uw p]
[z]	[z]	eggs	[eh g z]
[sh]	[ʃ]	squash	[s k w aa sh]
[zh]	[ʒ]	ambrosia	[ae m b r ow zh ax]
[ch]	[tʃ]	cherry	[ch eh r iy]
[jh]	[dʒ]	jar	[jh aa r]
[l]	[l]	licorice	[l ih k axr ix sh]
[w]	[w]	kiwi	[k iy w iy]
[r]	[r]	rice	[r ay s]
[y]	[j]	yellow	[y eh l ow]
[h]	[h]	honey	[h ah n iy]
Less commonly used phones and allophones			
[q]	[ʔ]	uh-oh	[q ah q ow]
[dx]	[ɾ]	butter	[b ah dx axr]
[nx]	[ɾ̃]	winner	[w ih nx axr]
[el]	[l̩]	table	[t ey b el]

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[iy]	[i]	lily	[l ih l iy]
[ih]	[ɪ]	lily	[l ih l iy]
[ey]	[eɪ]	daisy	[d ey z iy]
[eh]	[ɛ]	pen	[p eh n]
[ae]	[æ]	aster	[ae s t axr]
[aa]	[ɑ]	poppy	[p aa p iy]
[ao]	[ɔ]	orchid	[ao r k ix d]
[uh]	[ʊ]	wood	[w uh d]
[ow]	[oʊ]	lotus	[l ow dx ax s]
[uw]	[u]	tulip	[t uw l ix p]
[ah]	[ʌ]	buttercup	[b ah dx axr k ah p]
[er]	[ɝ]	bird	[b er d]
[ay]	[aɪ]	iris	[ay r ix s]
[aw]	[aʊ]	sunflower	[s ah n f l aw axr]
[oy]	[oɪ]	soil	[s oy l]

Reduced and uncommon phones

[ax]	[ə]	lotus	[l ow dx ax s]
[axr]	[ɚ]	heather	[h eh dh axr]
[ix]	[ɨ]	tulip	[t uw l ix p]
[ux]	[ʉ]	dude[1]	[d ux d]

The Speech Chain (Denes and Pinson)

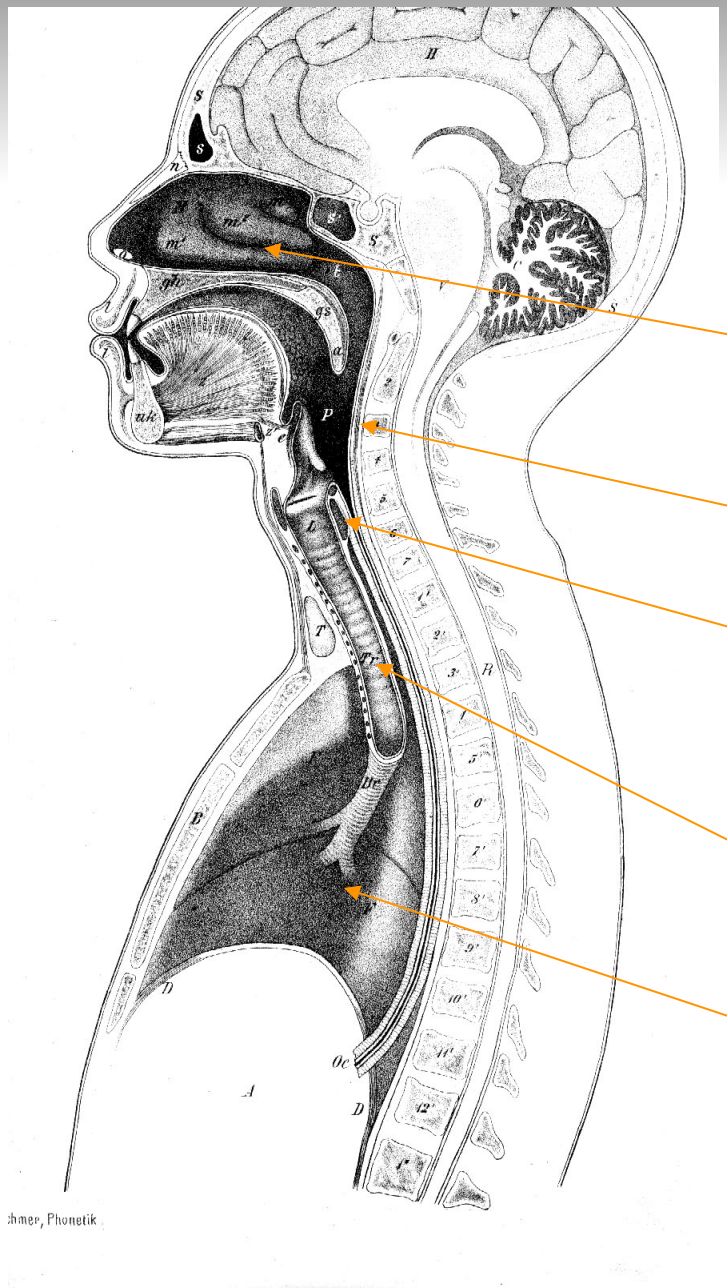


Speech Production Process

- Respiration:
  - We (normally) speak while breathing out. Respiration provides airflow. "Pulmonic egressive airstream"
- Phonation
  - Airstream sets vocal folds in motion. Vibration of vocal folds produces sounds. Sound is then modulated by:
- Articulation and Resonance
  - Shape of vocal tract, characterized by:
  - Oral tract
    - Teeth, soft palate (velum), hard palate
    - Tongue, lips, uvula
  - Nasal tract

Text adopted from Sharon Rose

Sagittal section of the vocal tract (Techmer 1880)



Nasal Cavity

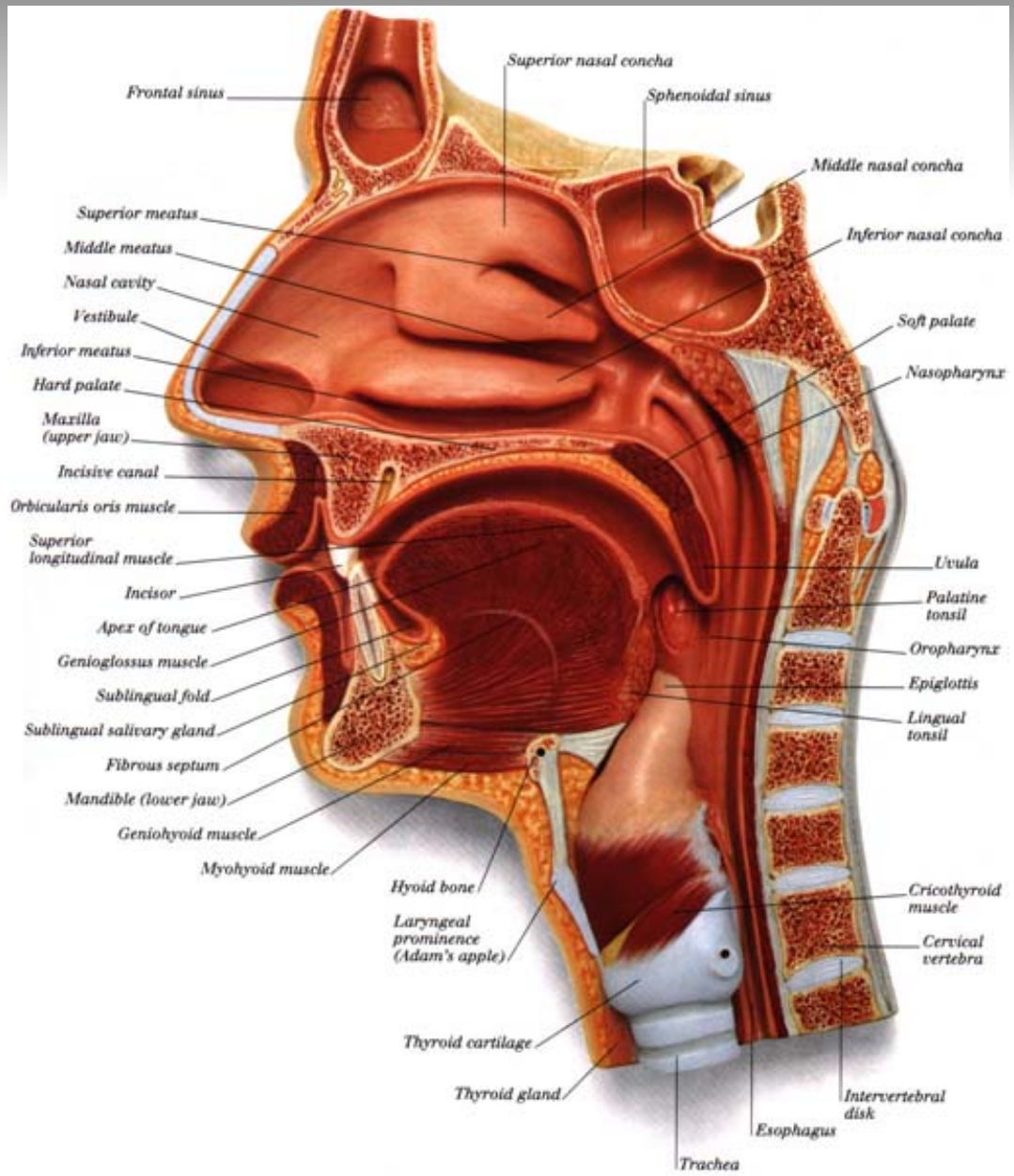
Pharynx

Vocal Folds (within the Larynx)

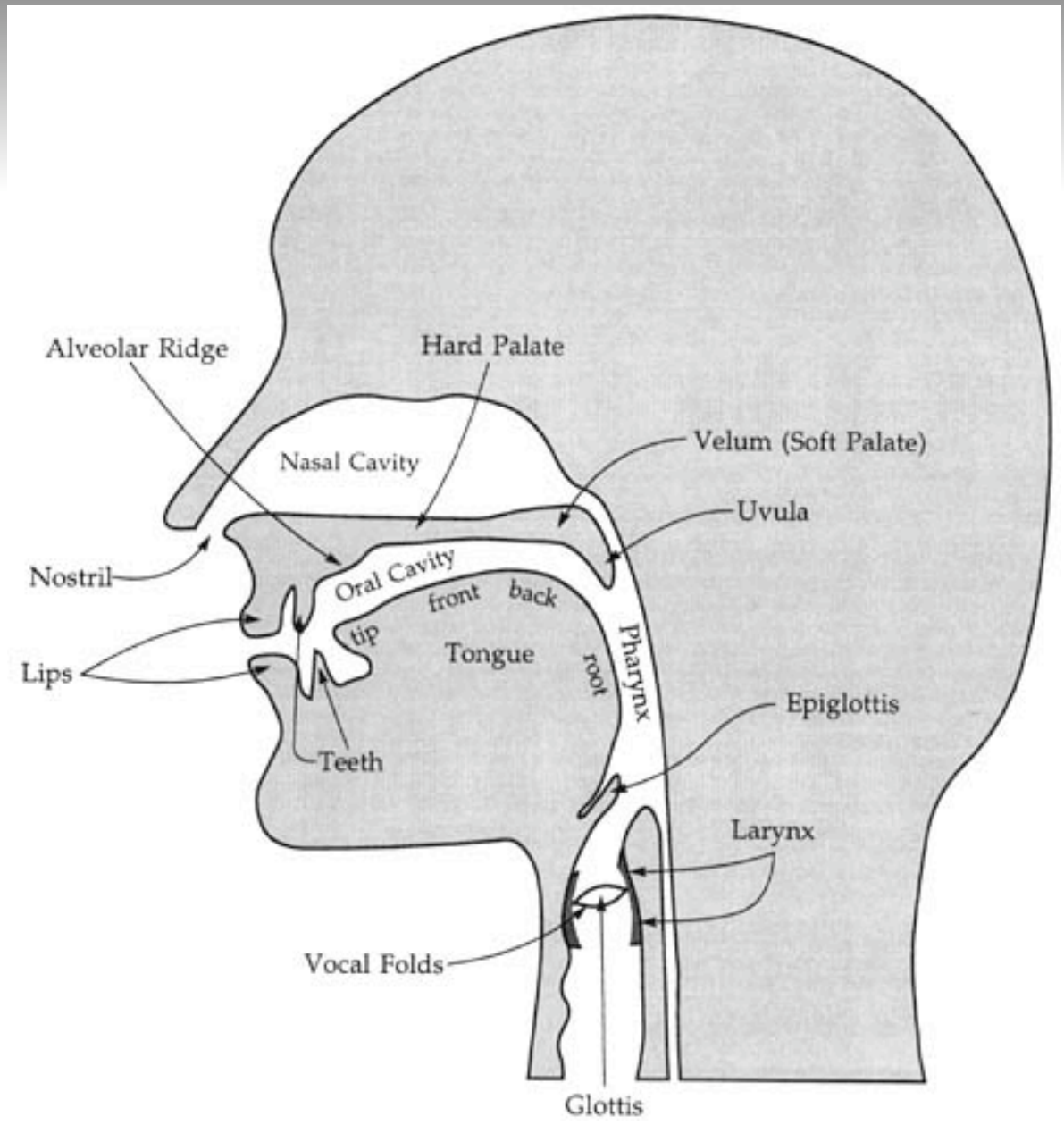
Trachea

Lungs

1/5/07



From Mark Liberman's website, from Ultimate Visual Dictionary



From Mark Liberman's Web Site, from Language Files (7th ed)

Vocal tract

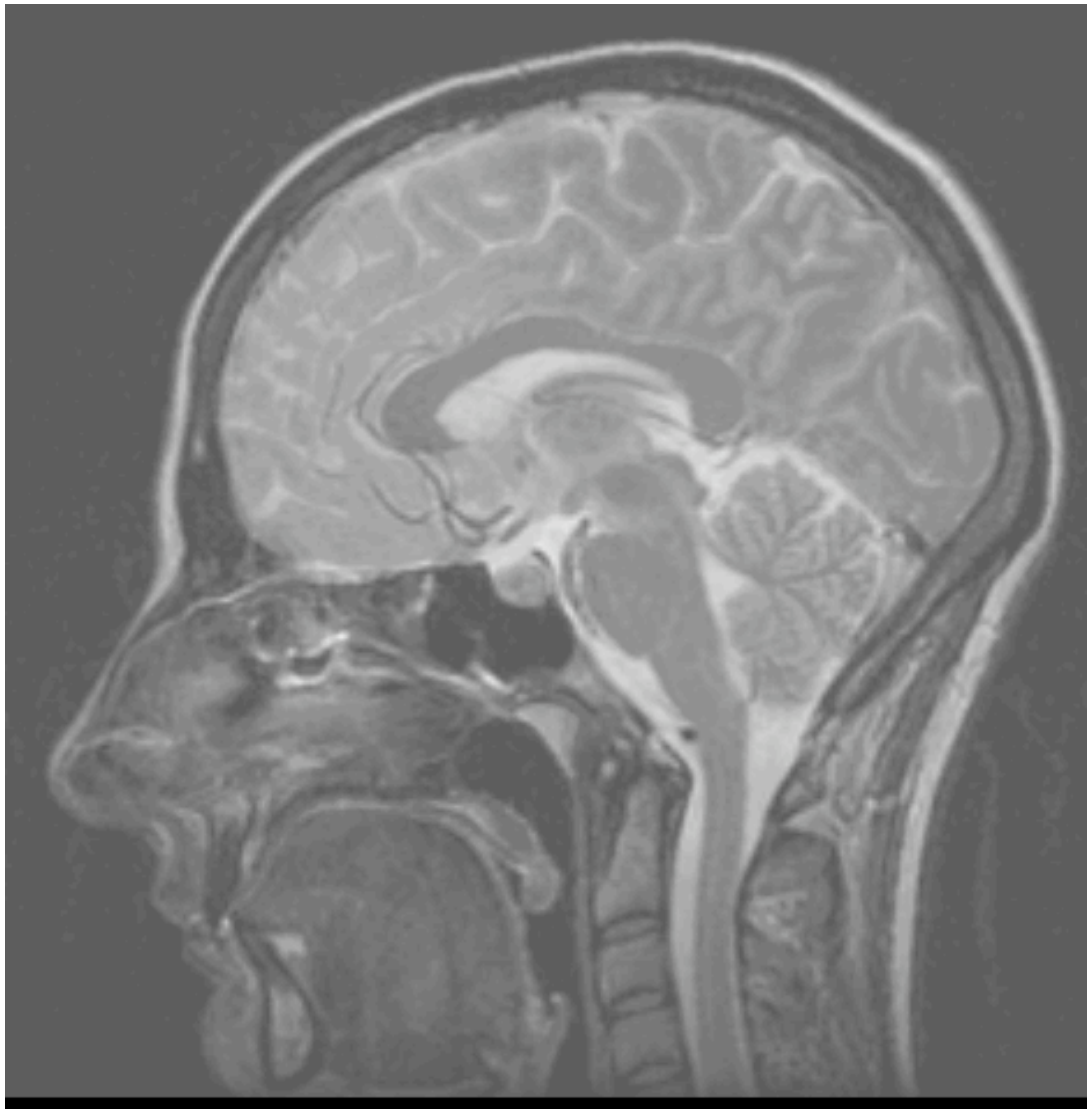


Figure thnx to John Coleman!!

Vocal tract movie (high speed x-ray)

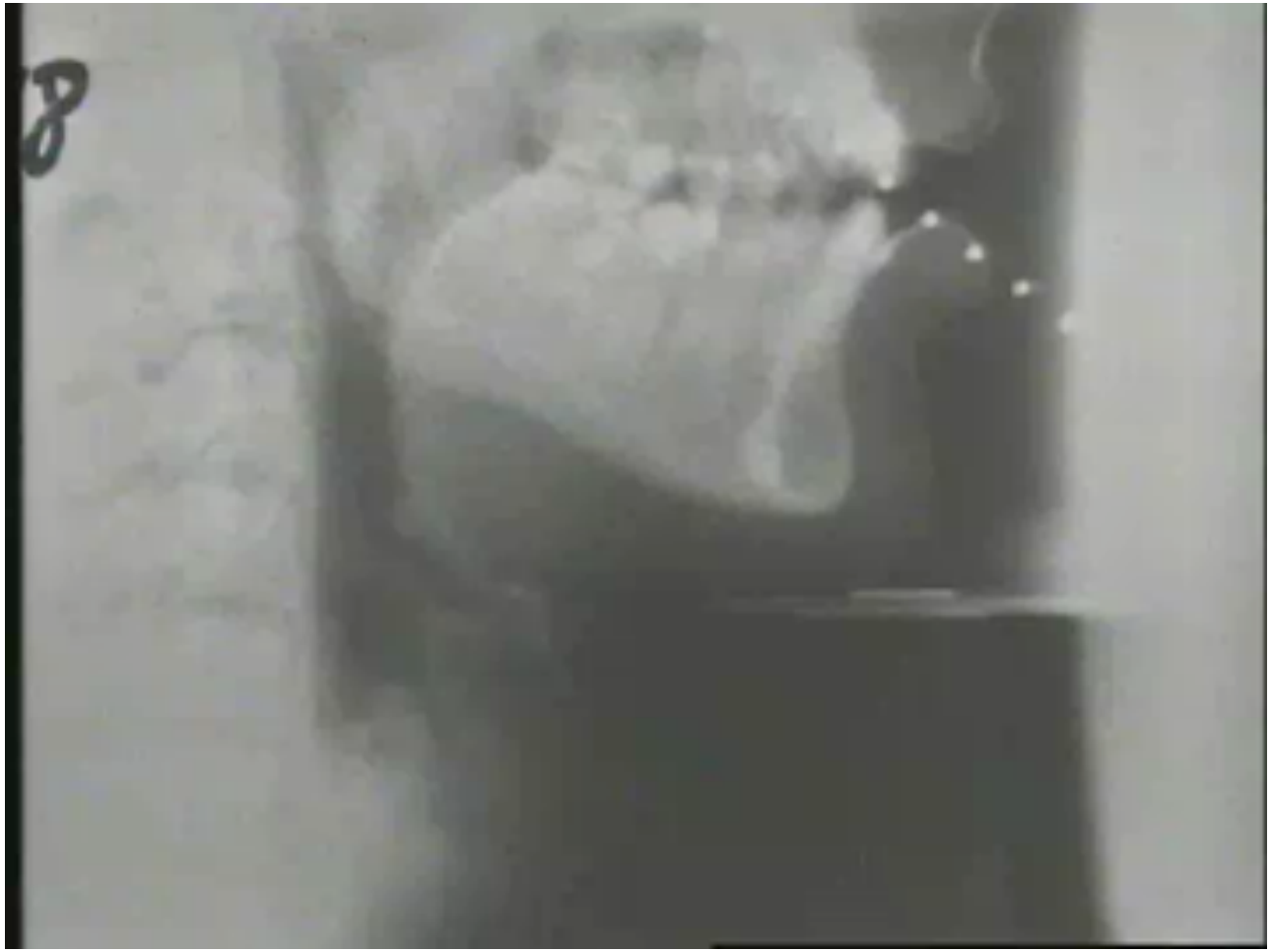


Figure of Ken Stevens, from Peter Ladefoged's web site

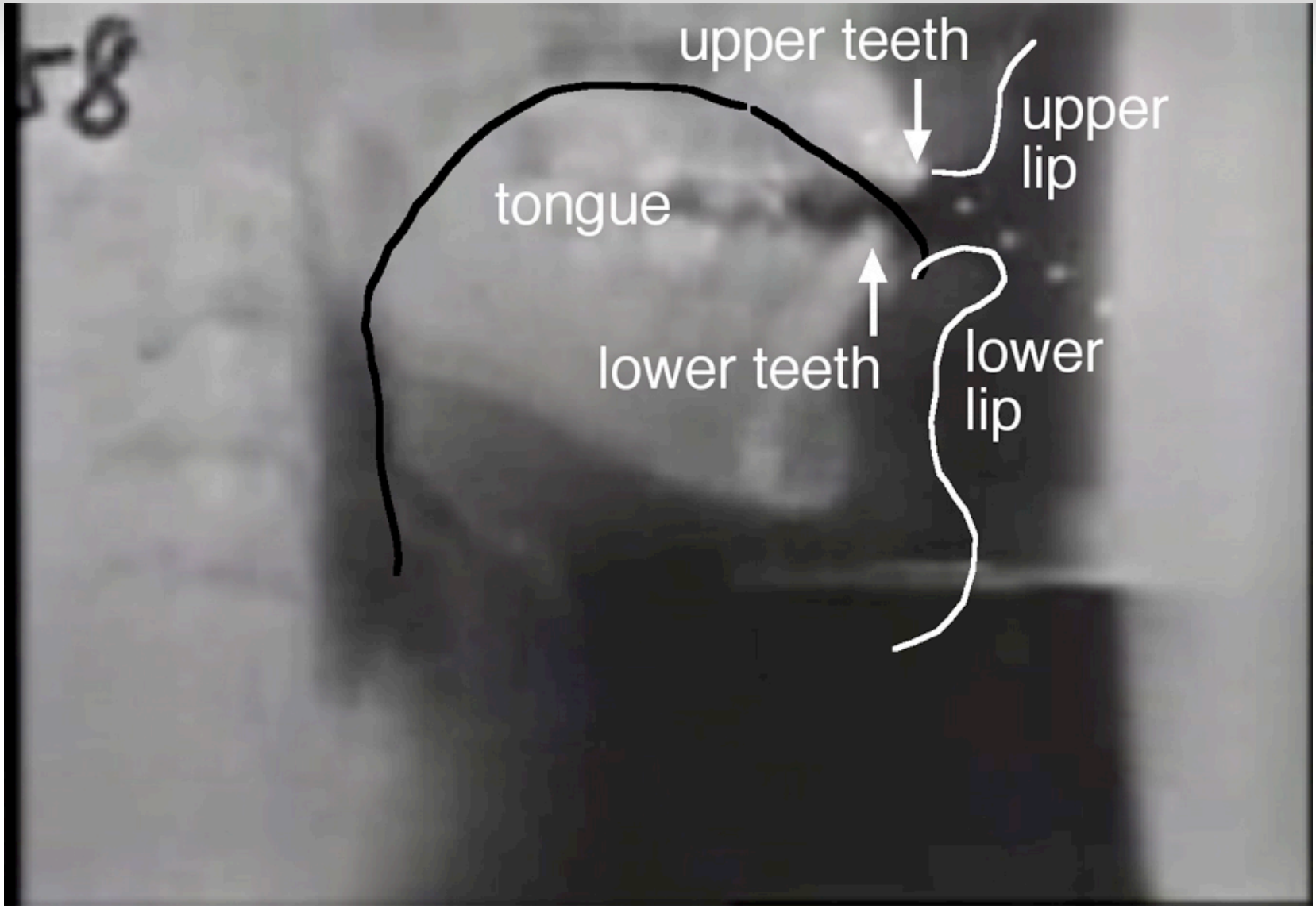


Figure of Ken Stevens, labels from Peter Ladefoged's web site

USC's SAIL Lab
Shri Narayanan

Larynx and Vocal Folds

- The Larynx (voice box)
  - A structure made of cartilage and muscle
  - Located above the trachea (windpipe) and below the pharynx (throat)
  - Contains the vocal folds
  - (adjective for larynx: laryngeal)
- Vocal Folds (older term: vocal cords)
  - Two bands of muscle and tissue in the larynx
  - Can be set in motion to produce sound (voicing)

The larynx, external structure, from front

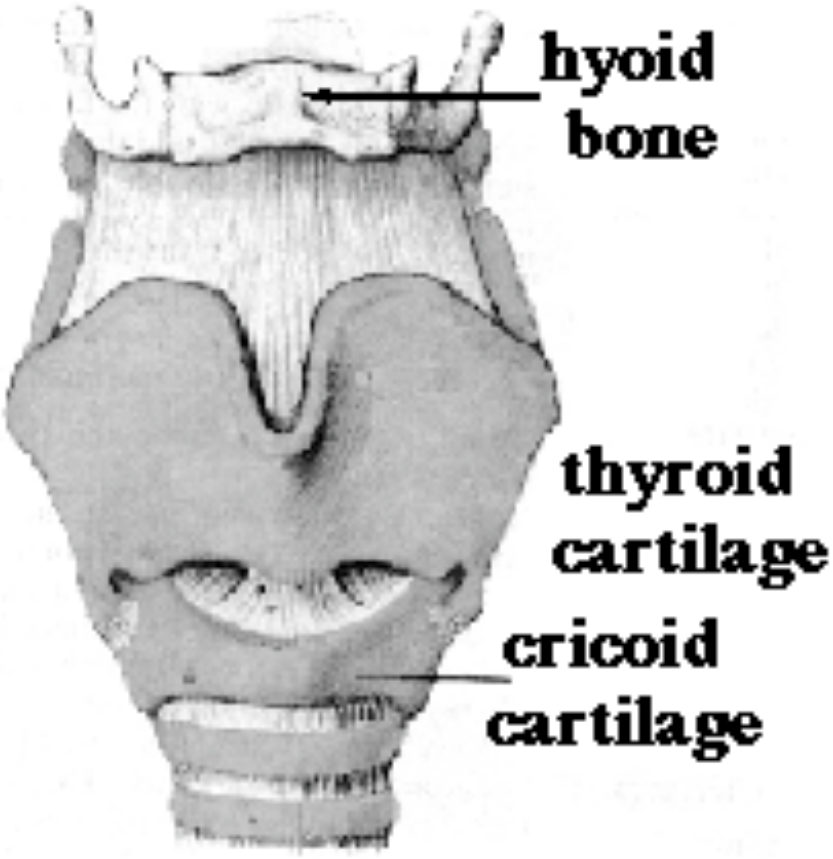


Figure thnx to John Coleman!!

Vertical slice through larynx, as seen from back

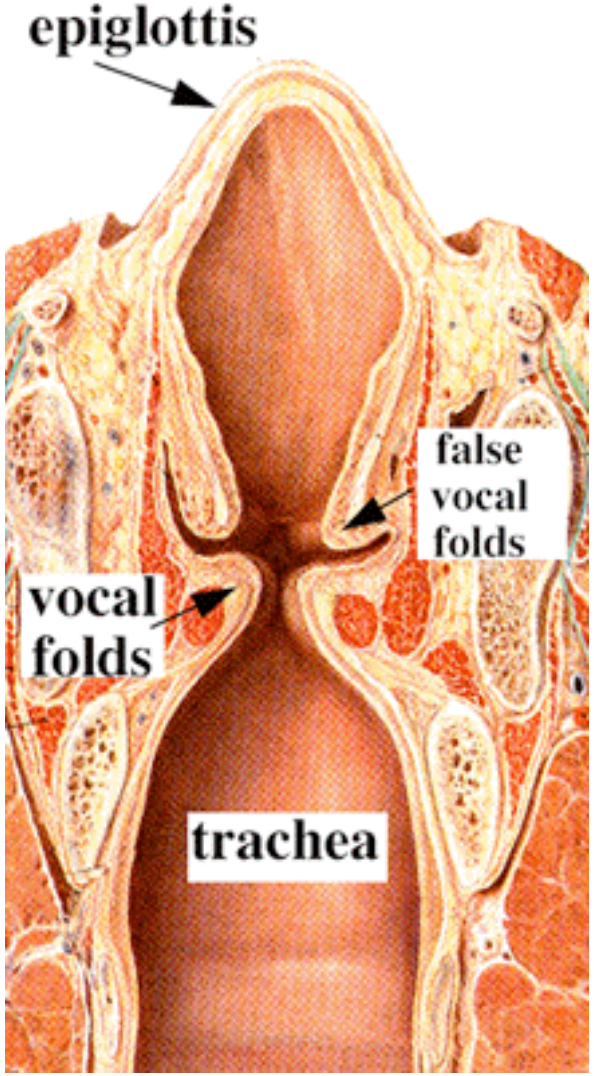
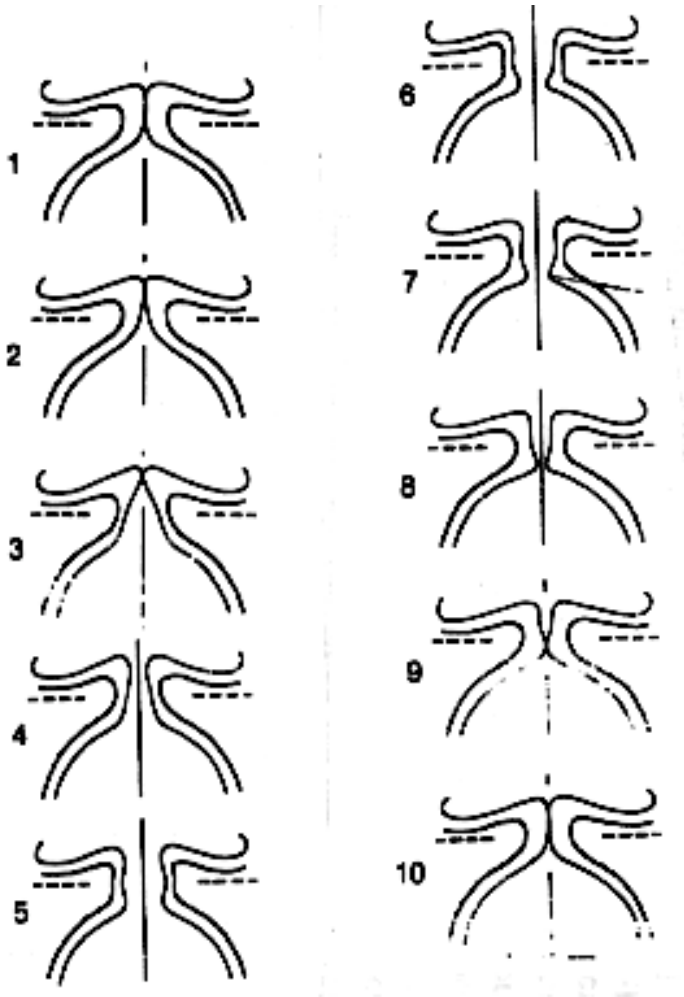


Figure thnx to John Coleman!!

Voicing:



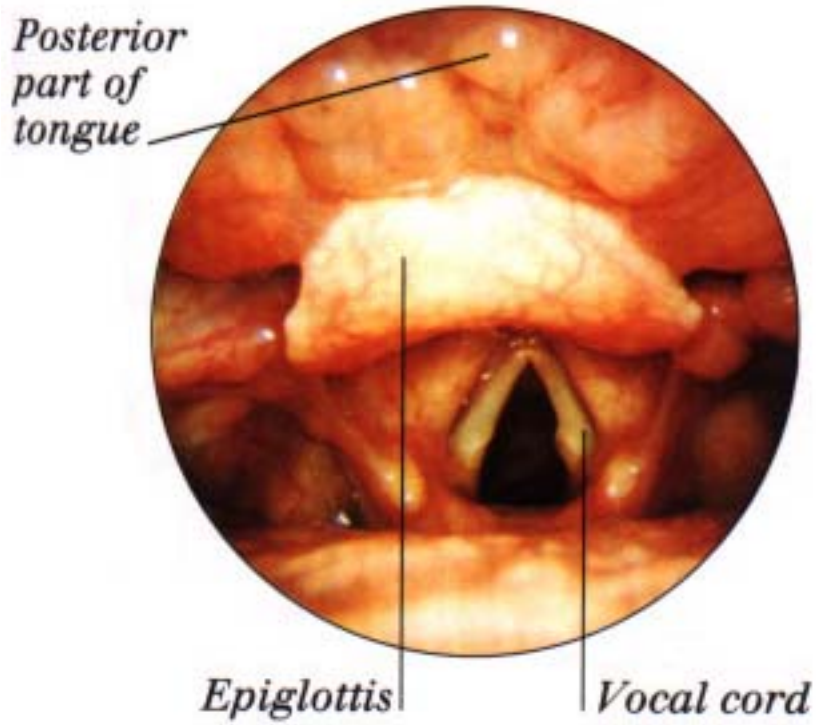
- Air comes up from lungs
- Forces its way through vocal cords, pushing open (2,3,4)
- This causes air pressure in glottis to fall, since:
  - when gas runs through constricted passage, its velocity increases (**Venturi tube effect**)
  - this increase in velocity results in a drop in pressure (**Bernoulli principle**)
- Because of drop in pressure, vocal cords snap together again (6-10)
- Single cycle: ~1/100 of a second.

Figure & text from John Coleman's web site

Voicelessness

- When vocal cords are open, air passes through unobstructed
- Voiceless sounds: p/t/k/s/f/sh/th/ch
- If the air moves very quickly, the turbulence causes a different kind of phonation: whisper

Vocal folds open during breathing



From Mark Liberman's web site, from Ultimate Visual Dictionary

Vocal Fold Vibration

UCLA Phonetics Lab Demo

Consonants and Vowels

- Consonants: phonetically, sounds with audible noise produced by a constriction
- Vowels: phonetically, sounds with no audible noise produced by a constriction

- (it's more complicated than this, since we have to consider syllabic function, but this will do for now)

Place of Articulation

- Consonants are classified according to the location where the airflow is most constricted.
- This is called place of articulation
- Three major kinds of place articulation:
  - Labial (with lips)
  - Coronal (using tip or blade of tongue)
  - Dorsal (using back of tongue)

Places of articulation

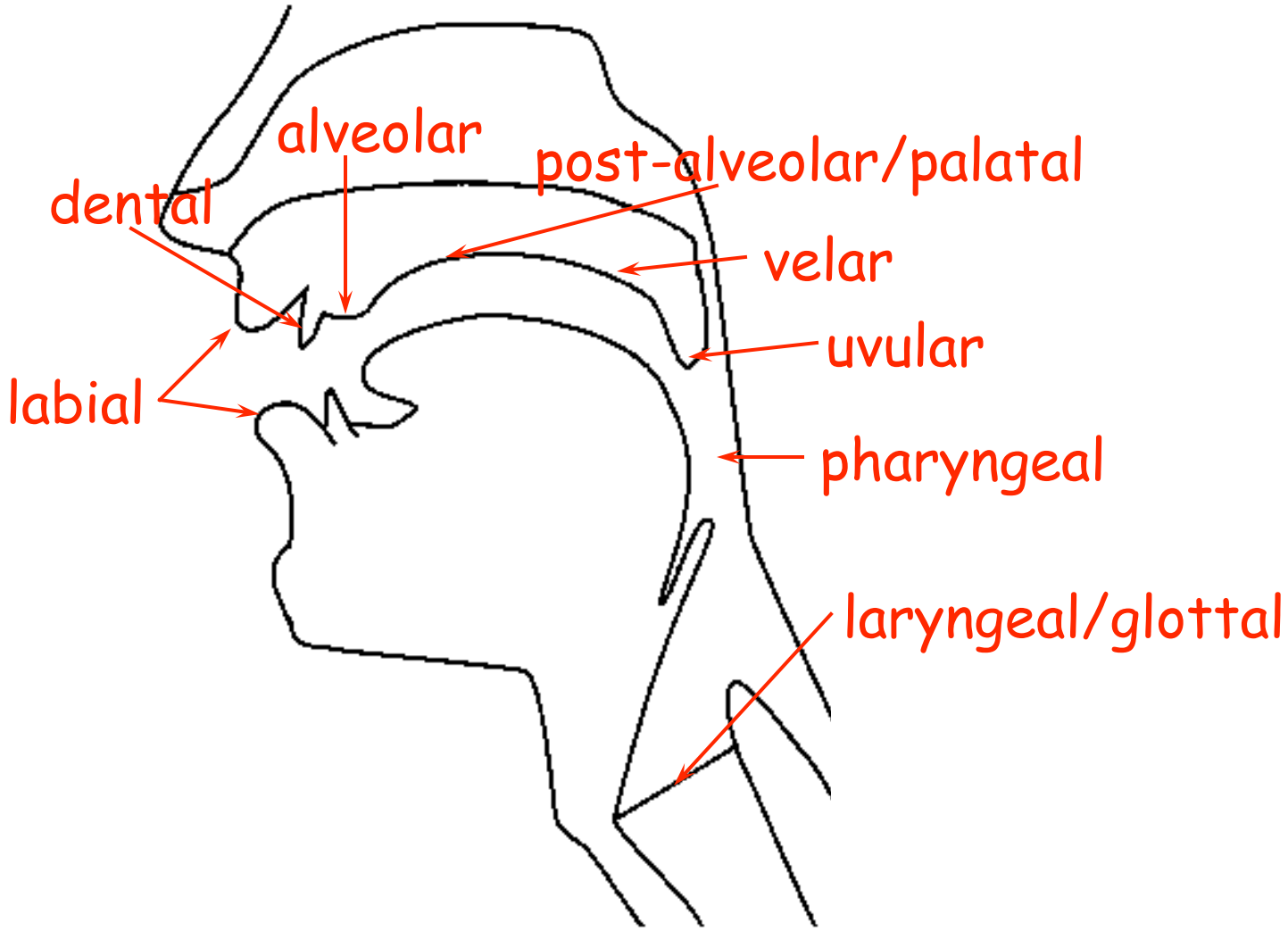
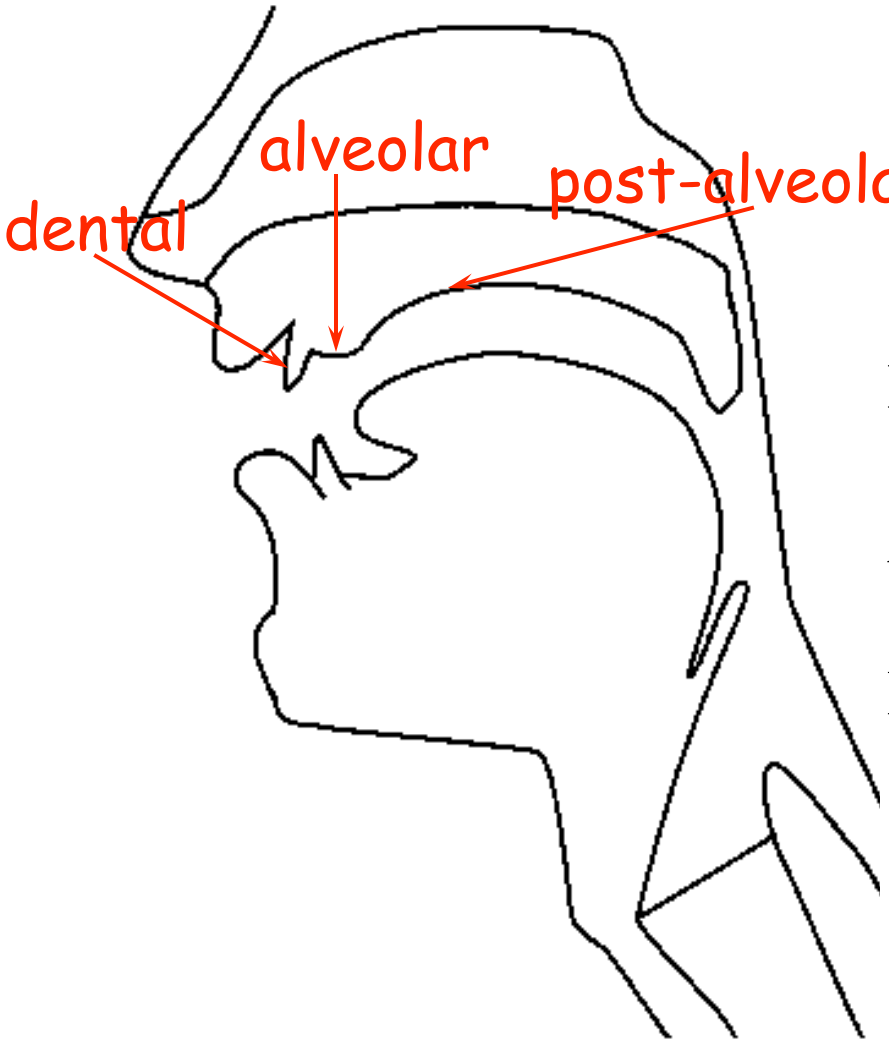


Figure thanks to Jennifer Venditti

Coronal place



Dental:
th/dh
Alveolar:
t/d/s/z/l
Post:
sh/zh/y

Figure thanks to Jennifer Venditti

Dorsal Place

Velar:

k/g/ng

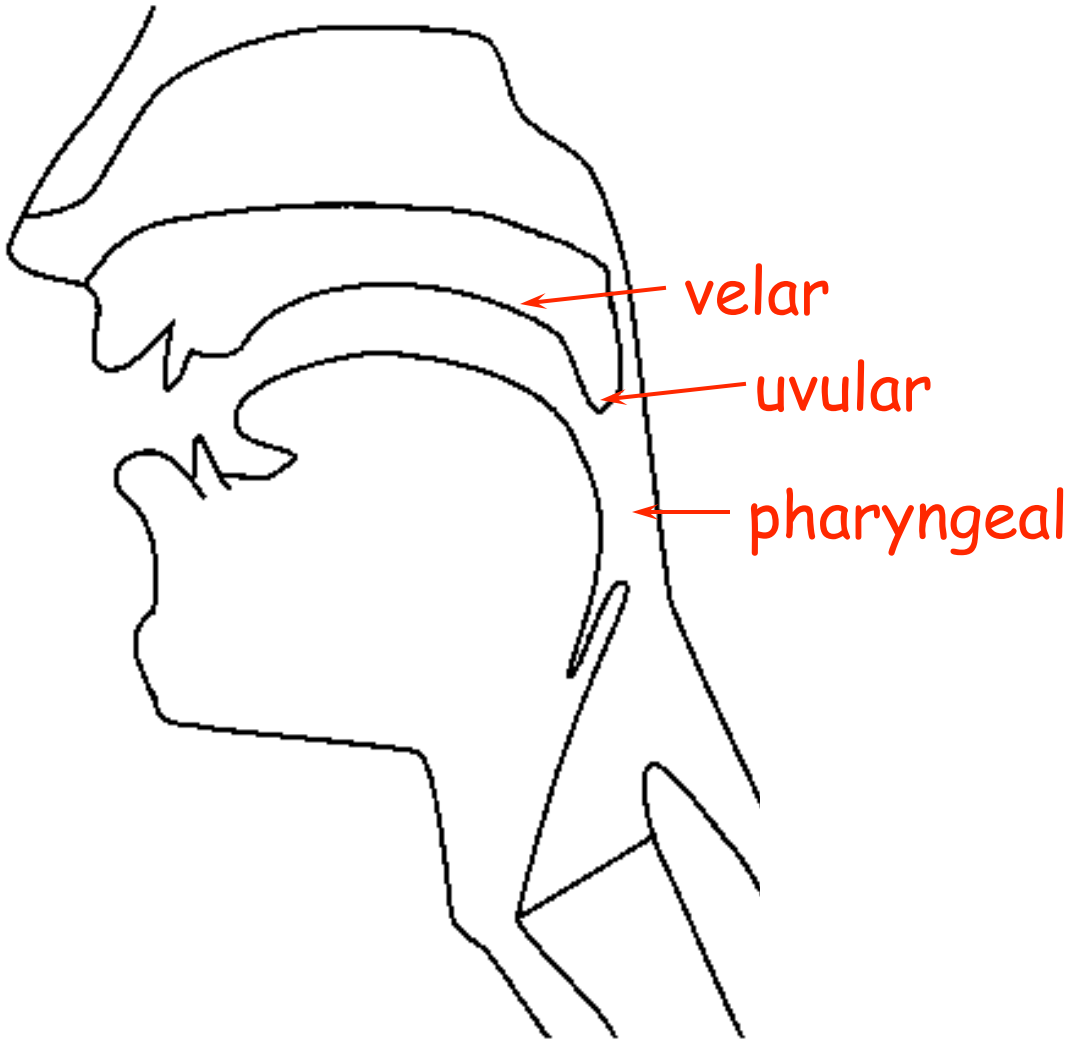
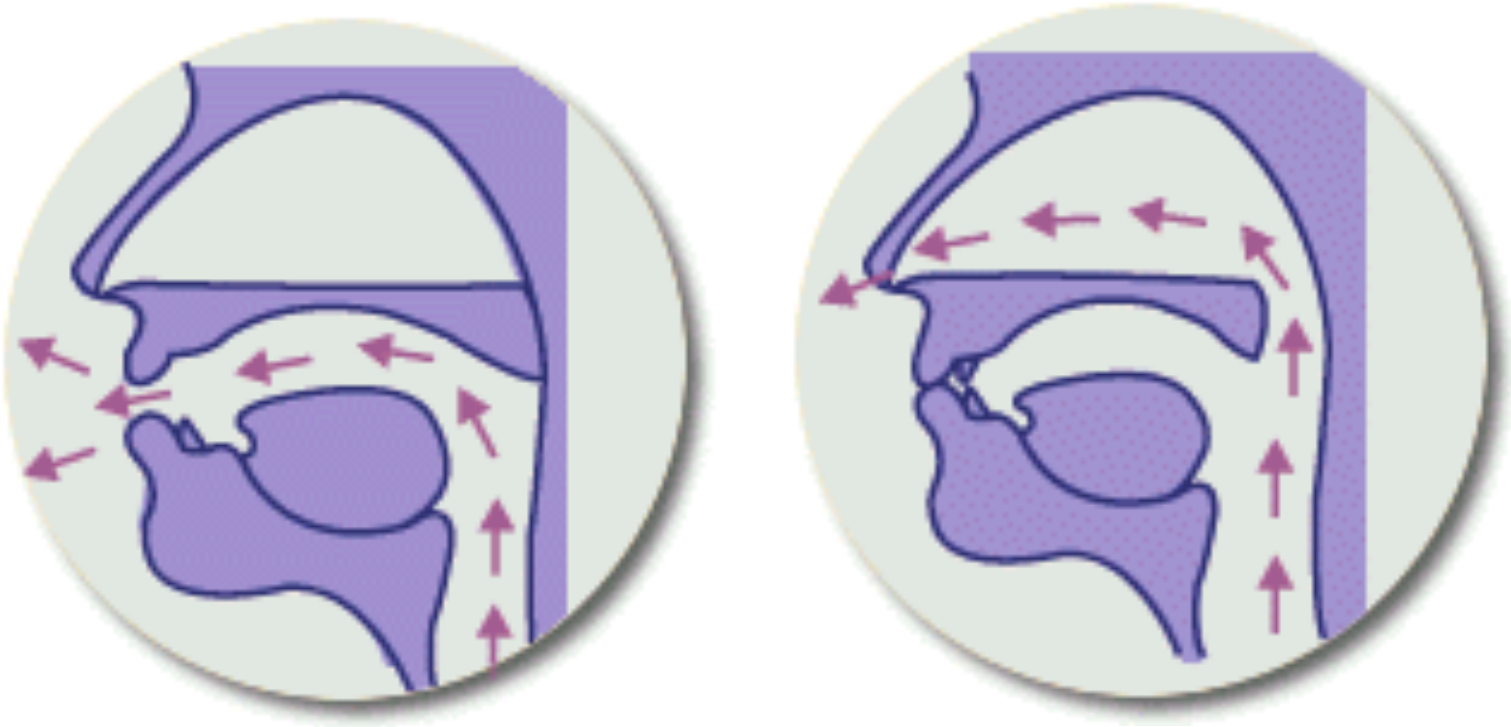


Figure thanks to Jennifer Venditti

Manner of Articulation

- Stop: complete closure of articulators, so no air escapes through mouth
- Oral stop: palate is raised, no air escapes through nose. Air pressure builds up behind closure, explodes when released
  - p, t, k, b, d, g
- Nasal stop: oral closure, but palate is lowered, air escapes through nose.
  - m, n, ng

Oral vs. Nasal Sounds



Thanks to Jong-bok Kim for this figure!

More on Manner of articulation of consonants

- Fricatives
  - Close approximation of two articulators, resulting in turbulent airflow between them, producing a hissing sound.
    - f, v, s, z, th, dh
- Approximant
  - Not quite-so-close approximation of two articulators, so no turbulence
    - y, r
- Lateral approximant
  - Obstruction of airstream along center of oral tract, with opening around sides of tongue.
    - l

Text from Ladefoged "A Course in Phonetics"

More on manner of articulation of consonants

- Tap or flap
  - Tongue makes a single tap against the alveolar ridge
    - dx in "butter"
- Affricate
  - Stop immediately followed by a fricative
    - ch, jh

Articulatory parameters for English consonants (in ARPAbet)

MANNER OF ARTICULATION	PLACE OF ARTICULATION															
		bilabial		labio-dental		inter-dental		alveolar		palatal		velar		glottal	
	stop	p	b					t	d			k	g	q	
	fric.			f	v	th	dh	s	z	sh	zh			h	
	affric.									ch	jh				
	nasal		m						n				ng		
	approx		w						l/r		y				
	flap							dx							

VOICING: voiceless | voiced

Table from Jennifer Venditti

Tongue position for vowels

Vowels

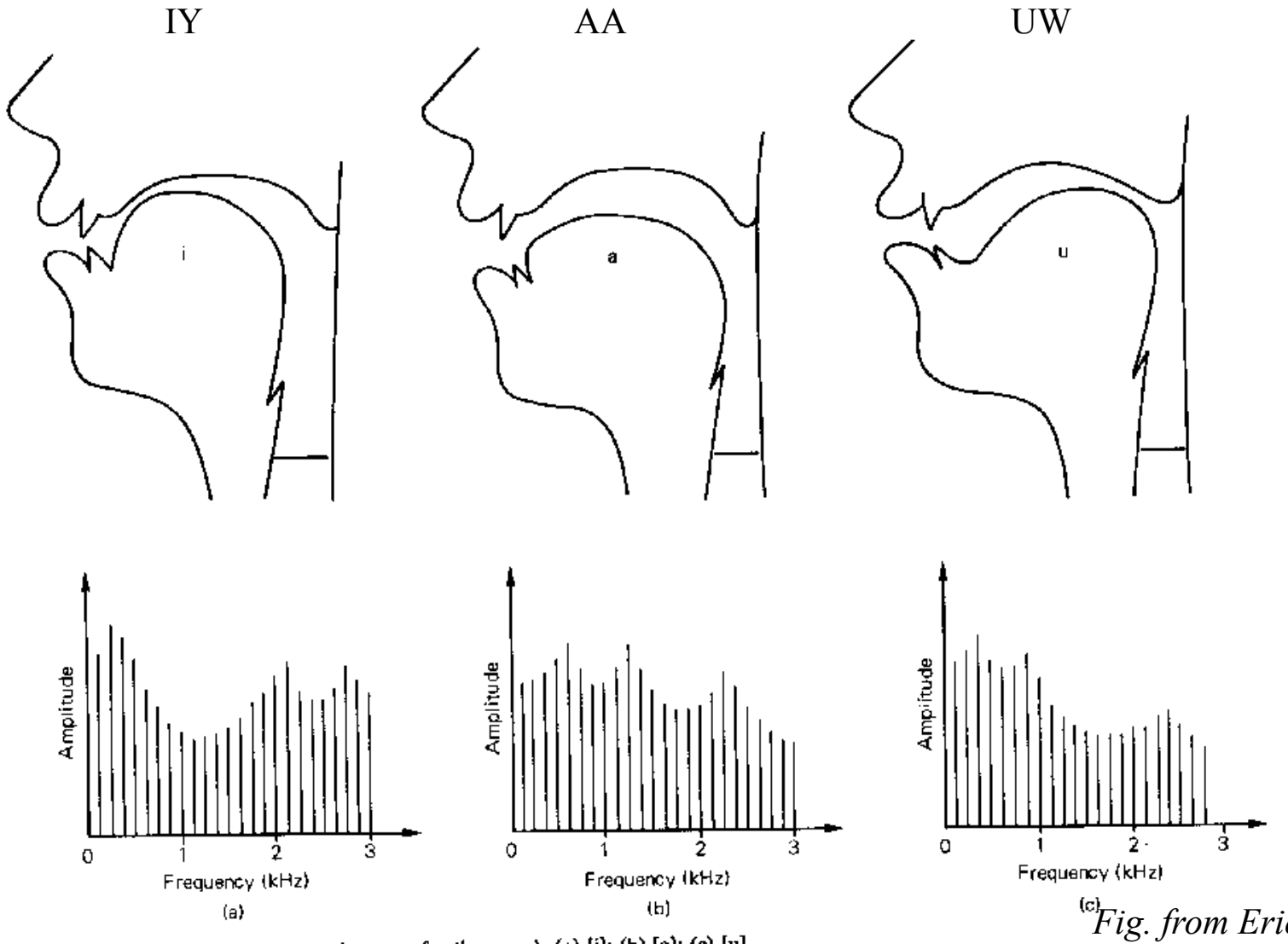


Fig. from Eric Keller

American English Vowel Space

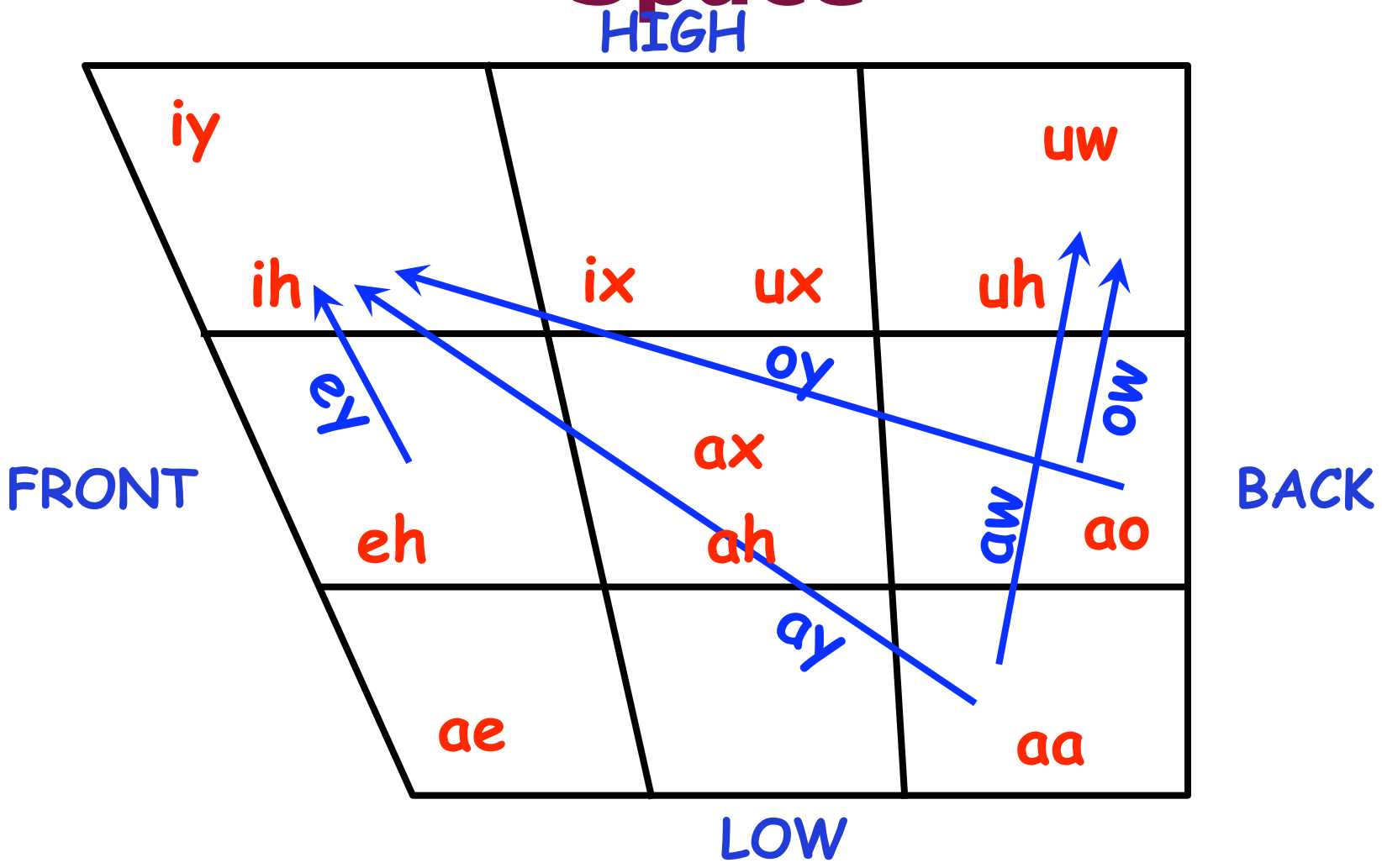


Figure from Jennifer Venditti

[iy] vs. [uw]

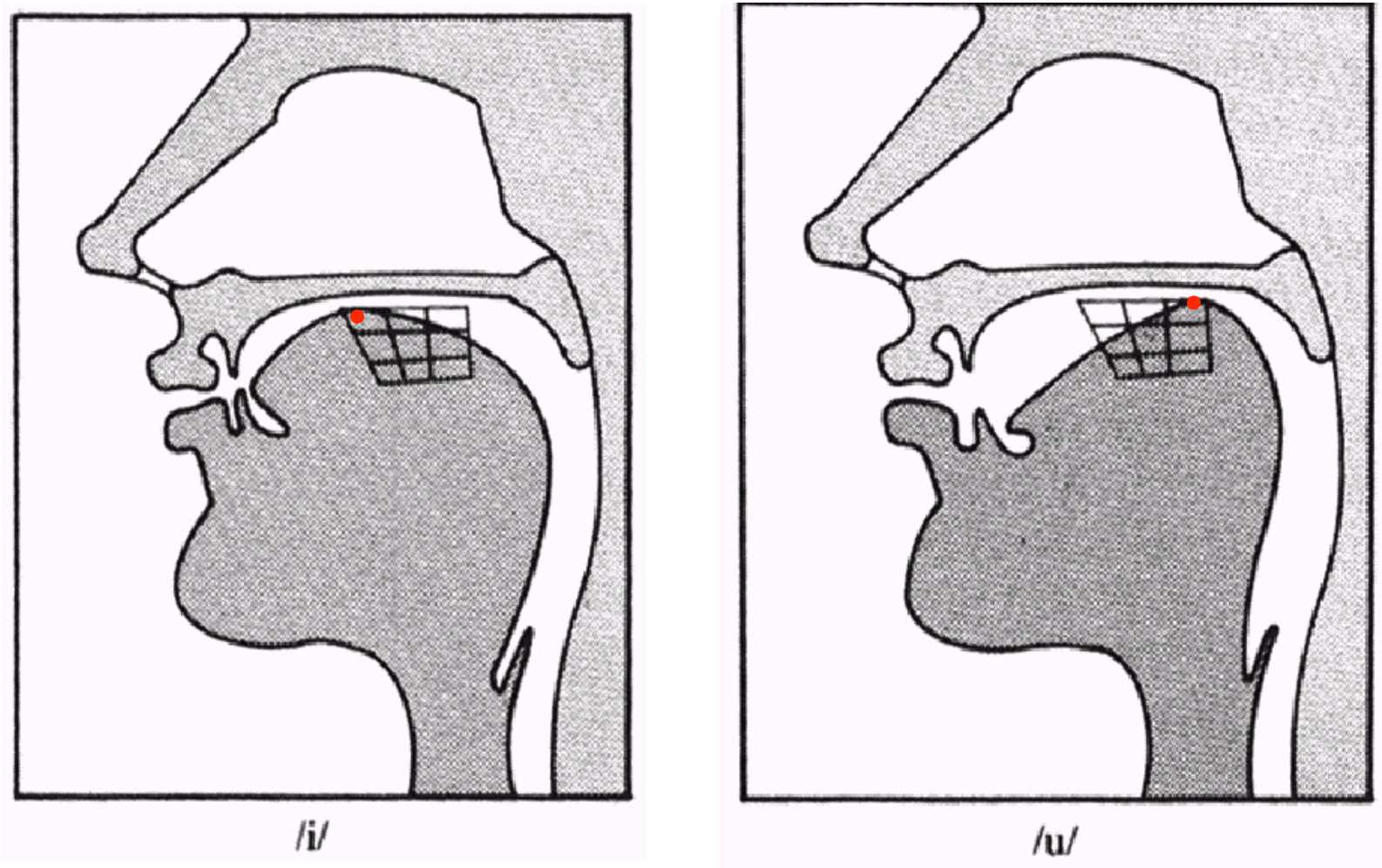


Figure from Jennifer Venditti, from a lecture given by Rochelle Newman

[ae] vs. [aa]

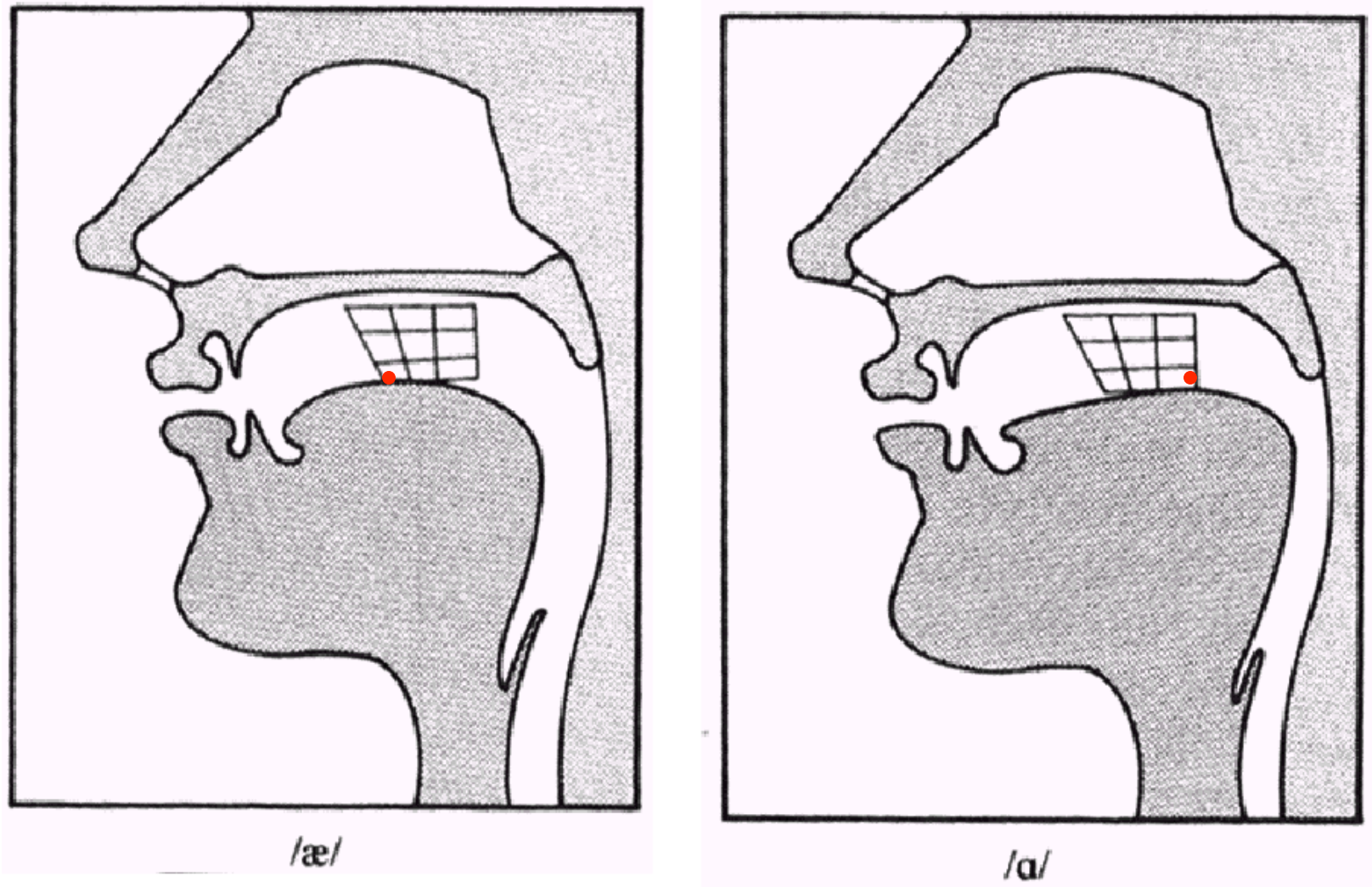
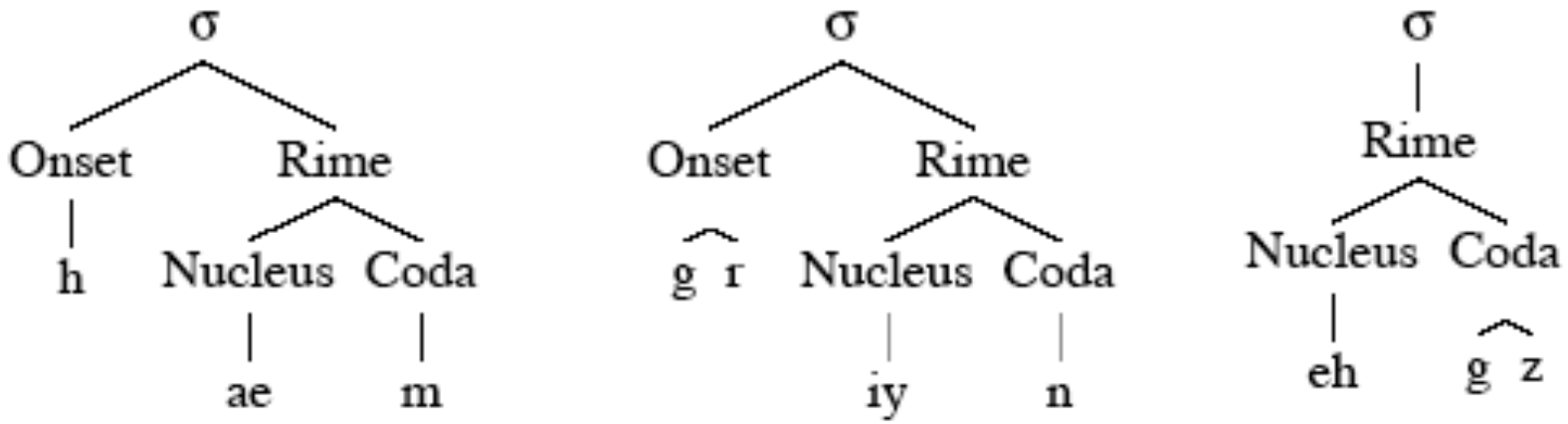


Figure from Jennifer Venditti, from a lecture given by Rochelle Newman

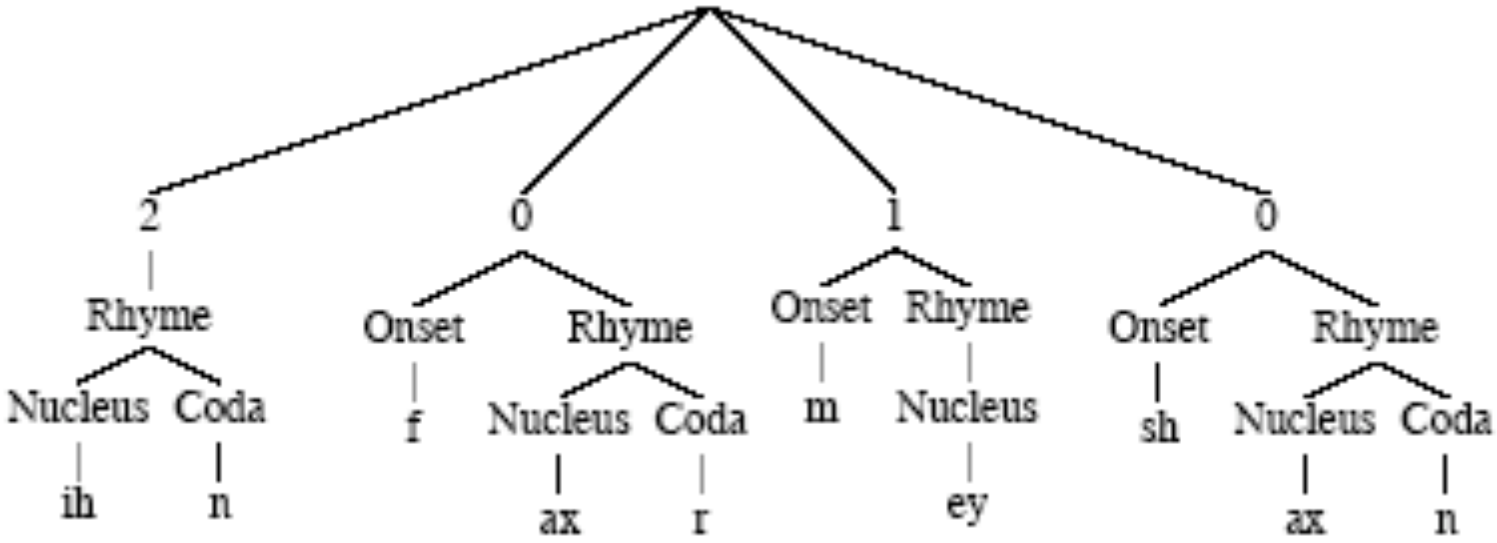
More phonetic structure

- Syllables
  - Composed of vowels and consonants. Not well defined. Something like a "vowel nucleus with some of its surrounding consonants".



More phonetic structure

- Stress
  - Some syllables have more energy than others
  - Stressed syllables versus unstressed syllables
  - (an) ‘INsult vs. (to) in’SULT
  - (an) ‘OBject vs. (to) ob’JECT
- Simple model: every multi-syllabic word has one syllable with:
  - **“primary stress”**
    - We can represent by using the number “1” on the vowel (and an implicit unmarking on the other vowels)
    - “table”: t ey1 b ax l
    - “machine: m ax sh iy1 n
  - Also possible: “secondary stress”, marked with a “2”
    - ih-2 n f axr m ey-1 sh ax n
  - Third category: **reduced**: schwa:
    - ax


2
0
1
0
Rhyme
Onset
Rhyme
Onset
Rhyme
Onset
Rhyme
Nucleus
Coda
f
Nucleus
Coda
m
Nucleus
sh
Nucleus
Coda
ih
n
ax
r
ey
ax
n

Where to go for more info

- Ladefoged, Peter. 1993. A Course in Phonetics
- Mark Liberman's site
  - http://www.ling.upenn.edu/courses/Spring_2001/ling001/phonetics.html
- John Coleman's site
  - http://www.phon.ox.ac.uk/%7Ejcoleman/mst_mphil_phonetics_course_index.html

Summary

- Overview and very brief history
- Articulatory Phonetics
- Administration
  - Overview of course topics
  - Grading
- ARPAbet transcription
- NEXT TIME: Acoustic phonetics