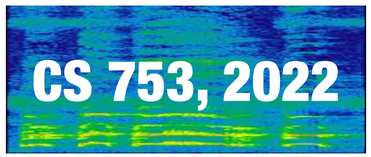
Live Session
(Basics of Speech Production)

Lecture 5a



Instructor: Preethi Jyothi, IITB

Cascaded ASR Systems: Putting it all together

- *A:* speech utterance
- O_A: acoustic features corresponding to the utterance *A*

$$W^* = \arg\max_W \Pr(O_A|W)\Pr(W)$$

- Return the word sequence that jointly assigns the highest probability to O_A
- How do we estimate Pr(O_A|W) and Pr(W)?
- How do we decode?

Acoustic Model

$$W^* = \arg\max_W \Pr(O_A|W) \Pr(W)$$

$$\Pr(O_A|W) = \sum_Q \Pr(O_A, Q|W)$$

$$= \sum_{q_1^T, w_1^N} \prod_{t=1}^{T} \Pr(O_t|O_1^{t-1}, q_1^t, w_1^N) \Pr(q_t|q_1^{t-1}, w_1^N)$$

First-order HMM assumptions

$$\approx \sum_{q_1^T, w_1^N} \prod_{t=1}^{T} \Pr(O_t|q_t, w_1^N) \Pr(q_t|q_{t-1}, w_1^N)$$

Viterbi approximation

$$\approx \max_{q_1^T, w_1^N} \prod_{t=1}^{T} \Pr(O_t|q_t, w_1^N) \Pr(q_t|q_{t-1}, w_1^N)$$

Acoustic Model

Transition probabilities

$$\Pr(O_A|W) = \max_{q_1^T, w_1^N} \prod_{t=1}^{T} \Pr(O_t|q_t, w_1^N) \Pr(q_t|q_{t-1}, w_1^N)$$

Emission probabilities

Modeled using a mixture of Gaussians

$$\Pr(O_t|q_t) = \sum_{m=1}^{M} c_{qm} \mathcal{N}(O|\mu_{qm}, \Sigma_{qm})$$

Language Model

$$W^* = \arg\max_W \Pr(O_A|W)\Pr(W)$$

$$\begin{aligned}\Pr(W) &= \Pr(w_1, w_2, \ldots, w_N) \\ &= \Pr(w_1)\ldots\Pr(w_N|w_{N-m+1}^{N-1})\end{aligned}$$

m-gram language model

- Further optimized using smoothing and interpolation with lower-order Ngram models

Decoding

$$W^* = \arg\max_{W} \Pr(O_A|W)\Pr(W)$$

$$W^* = \arg\max_{w_1^N, N} \left\{ \left[\prod_{n=1}^{N} \Pr(w_n|w_{n-m+1}^{n-1}) \right] \left[\sum_{q_1^T, w_1^N} \prod_{t=1}^{T} \Pr(O_t|q_t, w_1^N) \Pr(q_t|q_{t-1}, w_1^N) \right] \right\}$$

Viterbi

$$\approx \arg\max_{w_1^N, N} \left\{ \left[\prod_{n=1}^{N} \Pr(w_n|w_{n-m+1}^{n-1}) \right] \left[\max_{q_1^T, w_1^N} \prod_{t=1}^{T} \Pr(O_t|q_t, w_1^N) \Pr(q_t|q_{t-1}, w_1^N) \right] \right\}$$

- Search space still very huge for LVCSR tasks! Use approximate decoding techniques (A* decoding, beam-width decoding, etc.) to visit only promising parts of the search space

How are ASR systems evaluated?

- Word/Phone error rate (ER) uses the Levenshtein distance measure: What are the minimum number of edits (insertions/deletions/substitutions) required to convert W^* to $\mathrm{W}_{\mathrm{ref}}$?

On a test set with N instances:

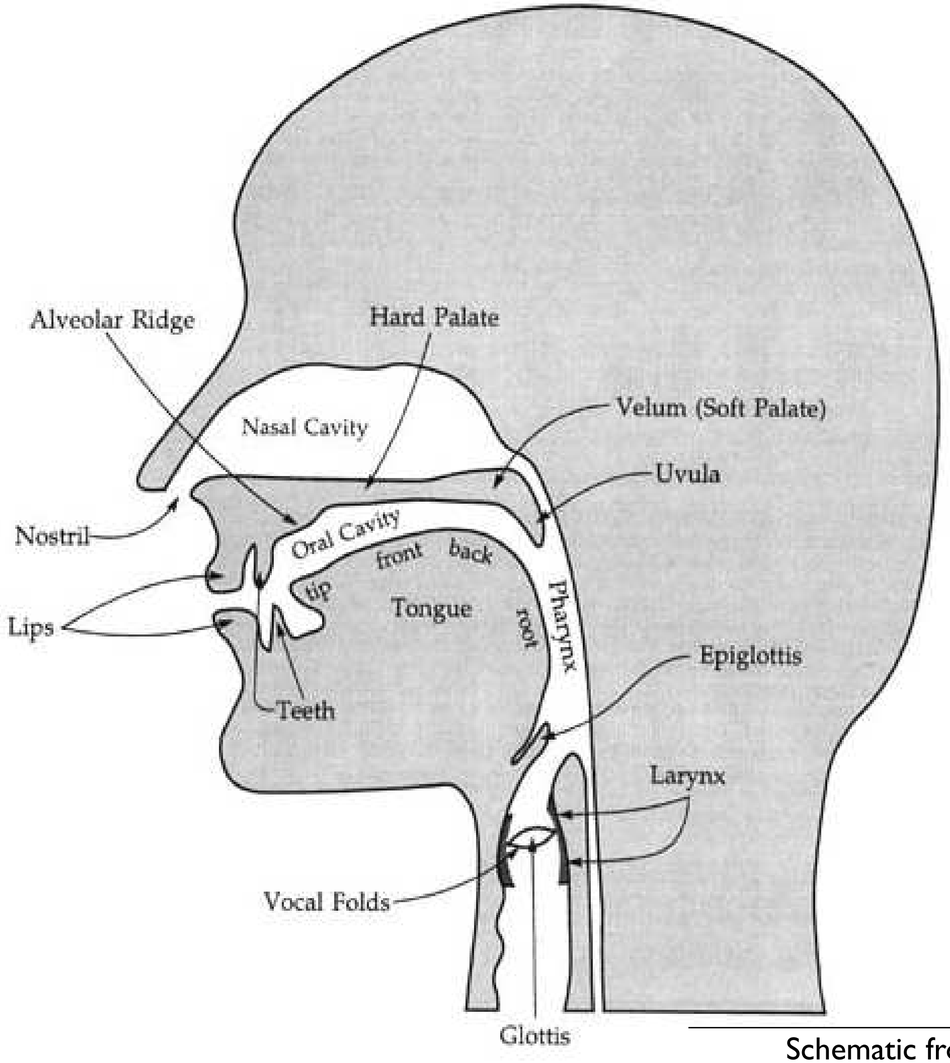
$$\mathrm{ER} = \frac{\sum_{j=1}^{N} \mathrm{Ins}_j + \mathrm{Del}_j + \mathrm{Sub}_j}{\sum_{j=1}^{N} \ell_j}$$

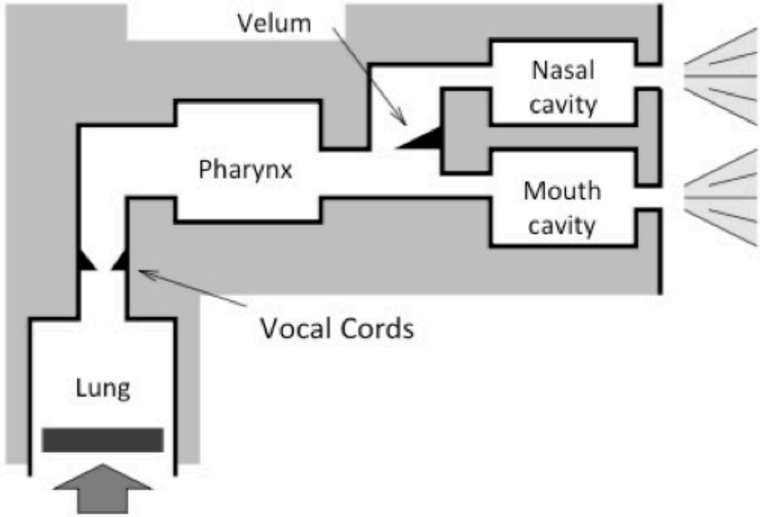
$\mathrm{Ins_j}$, $\mathrm{Del_j}$, $\mathrm{Sub_j}$ are number of insertions/deletions/substitutions in the j^{th} ASR output

ℓ_j is the total number of words/phones in the j^{th} reference

Basics of Speech Production

Speech Production

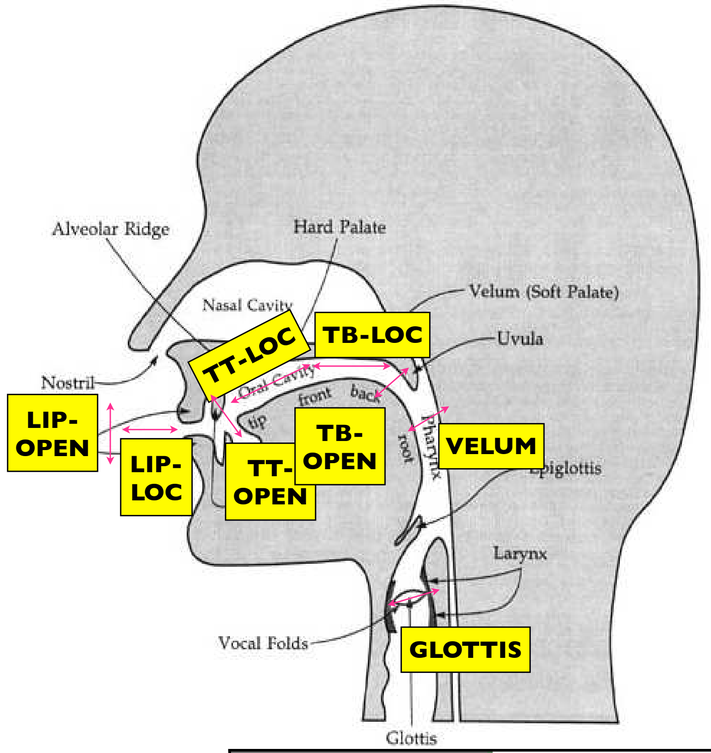




Schematic representation of the vocal organs

Schematic from L.Rabiner and B.-H.Juang , Fundamentals of speech recognition, 1993
Figure from http://www.phon.ucl.ac.uk/courses/spsci/iss/week6.php

Pronunciation Model



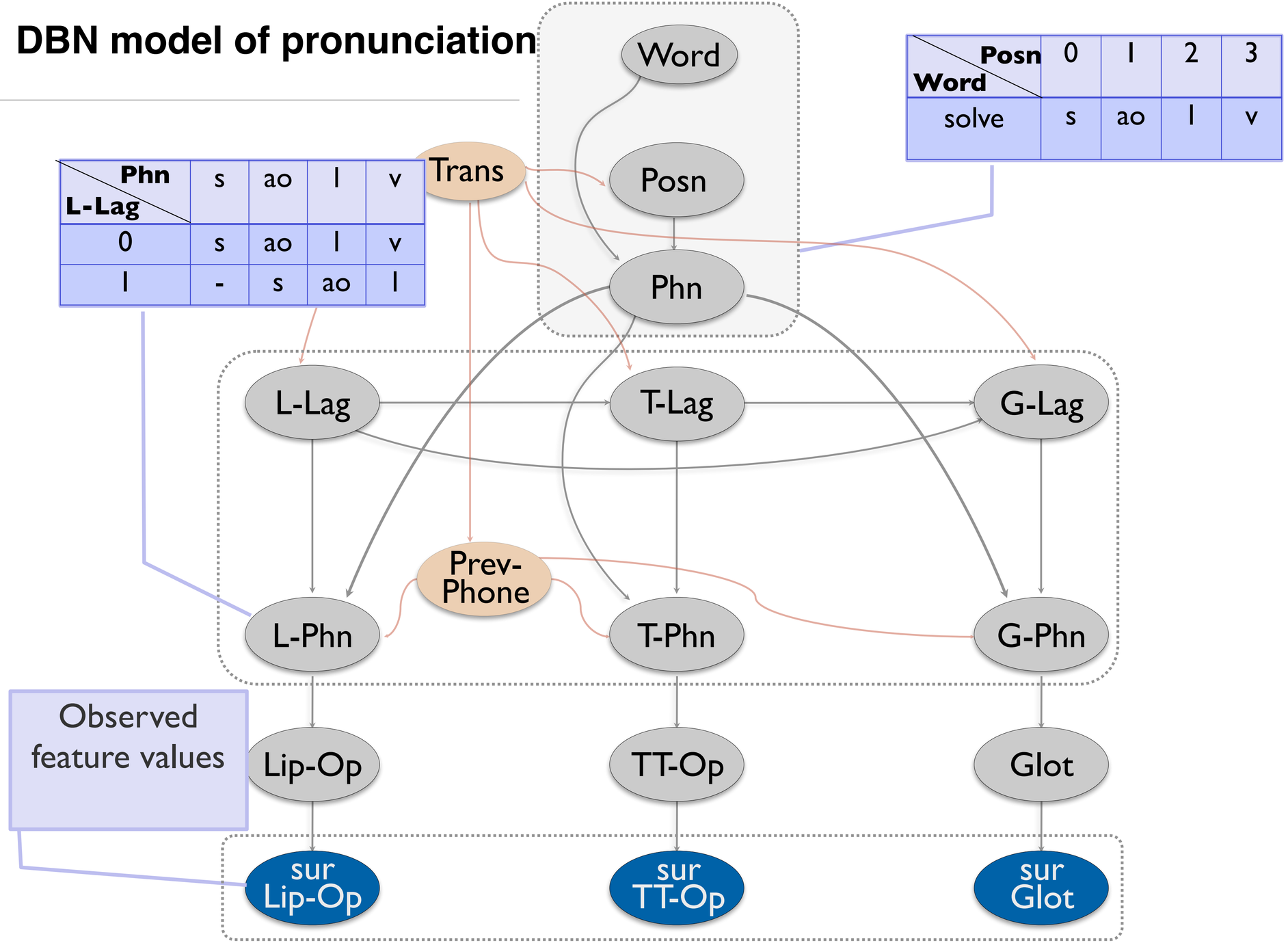
Articulatory Features

Parallel streams of articulator movements

Based on theory of articulatory phonology

PHONE	s	eh	n	s
LIP	open/labial			
TON.TIP	critical/alveolar	mid/alveolar	closed/alveolar	critical/alveolar
TON.BODY	mid/uvular	mid/palatal	mid/uvular	
GLOTTIS	open	critical		open
VELUM	closed		open	closed

DBN model of pronunciation

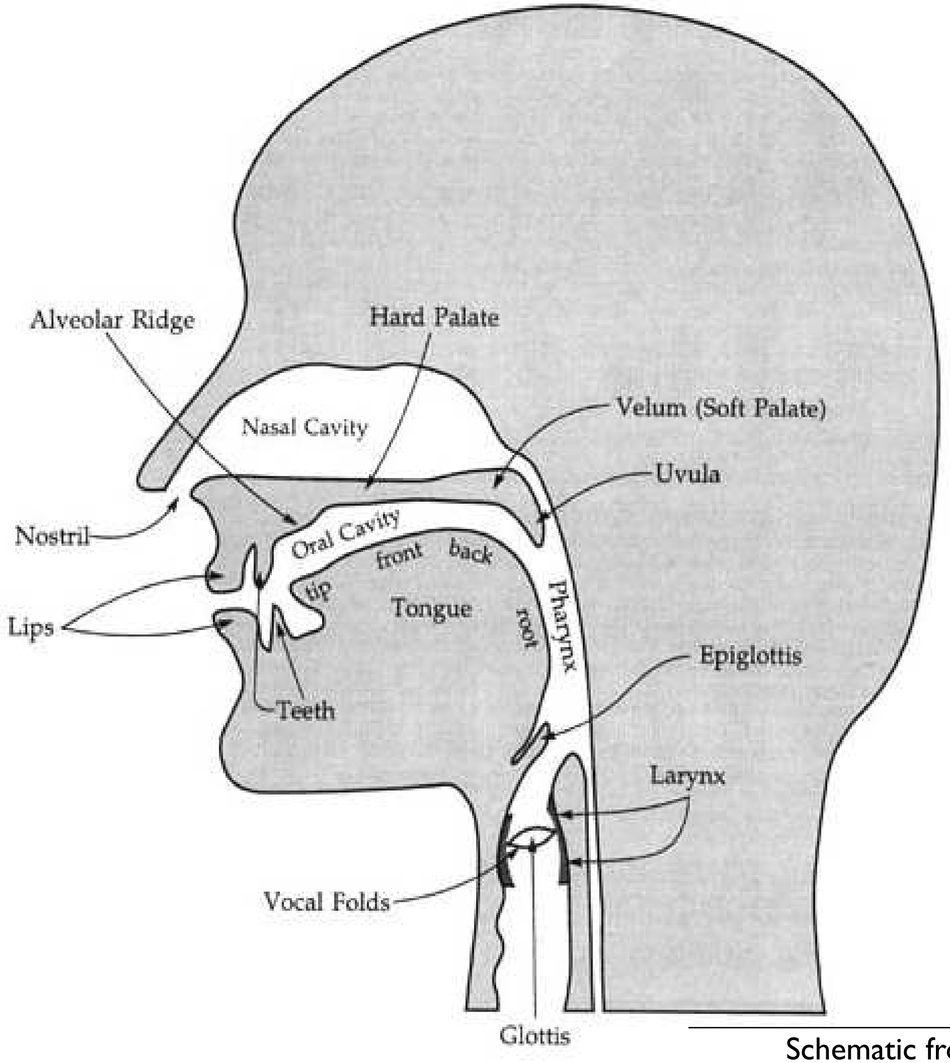


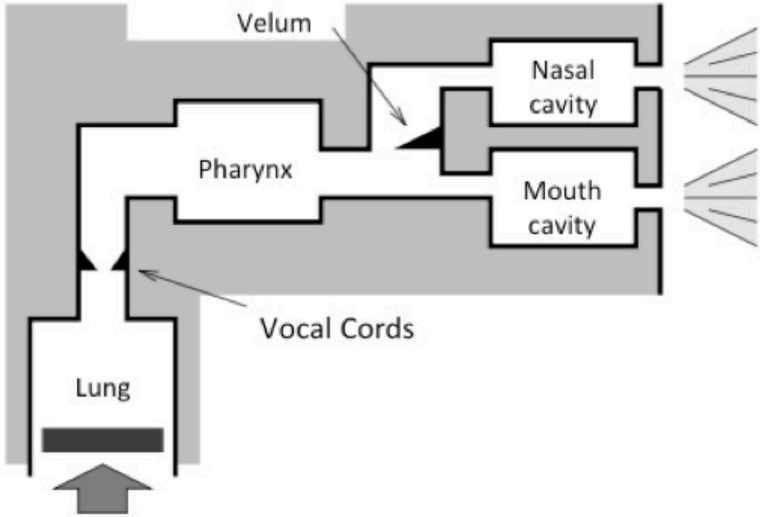
Word \ Posn	0	1	2	3
solve	s	ao	l	v

L-Lag \ Phn	s	ao	l	v
0	s	ao	l	v
1	-	s	ao	l

P. Jyothi, E. Fosler-Lussier & K. Livescu, IS'12. (Adapted from Livescu & Glass, Interspeech '04)

Speech Production





Schematic representation of the vocal organs

Schematic from L.Rabiner and B.-H.Juang , Fundamentals of speech recognition, 1993
Figure from http://www.phon.ucl.ac.uk/courses/spsci/iss/week6.php

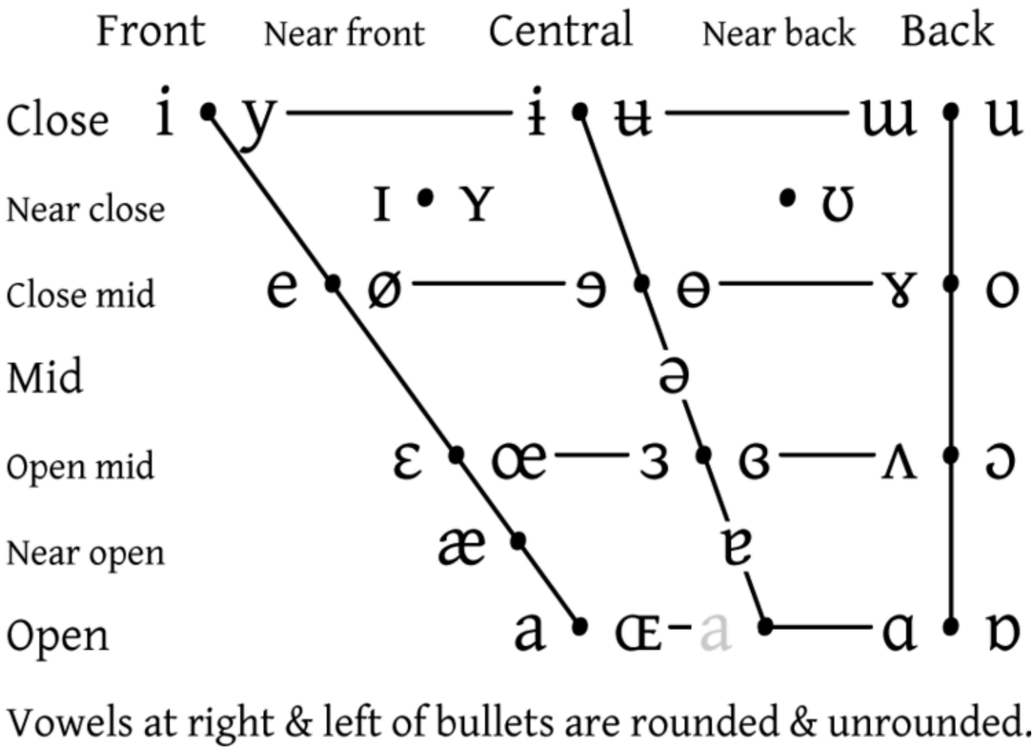
Sound units

- **Phones** are acoustically distinct units of speech
- **Phonemes** are abstract linguistic units that impart different meanings in a given language
  - Minimal pair: pan vs. ban
- **Allophones** are different acoustic realisations of the same phoneme
- **Phonetics** is the study of speech sounds and how they're produced
- **Phonology** is the study of patterns of sounds in different languages

Vowels

- Sounds produced with no obstruction to the flow of air through the vocal tract

VOWEL QUADRILATERAL



Vowels at right & left of bullets are rounded & unrounded.

Image from https://en.wikipedia.org/wiki/File:IPA_vowel_chart_2005.png

Spectrogram

- Spectrogram is a sequence of spectra stacked together in time, with amplitude of the frequency components expressed as a heat map

- Spectrograms of certain vowels: http://www.phon.ucl.ac.uk/courses/spsci/iss/week5.php

- Praat (http://www.fon.hum.uva.nl/praat/) is a good toolkit to analyse speech signals (plot spectrograms, generate pitch contours, etc.)

Consonants (voicing/place/manner)

- “Consonants are made by restricting or blocking the airflow in some way, and may be voiced or unvoiced.” (J&M, Ch. 7)

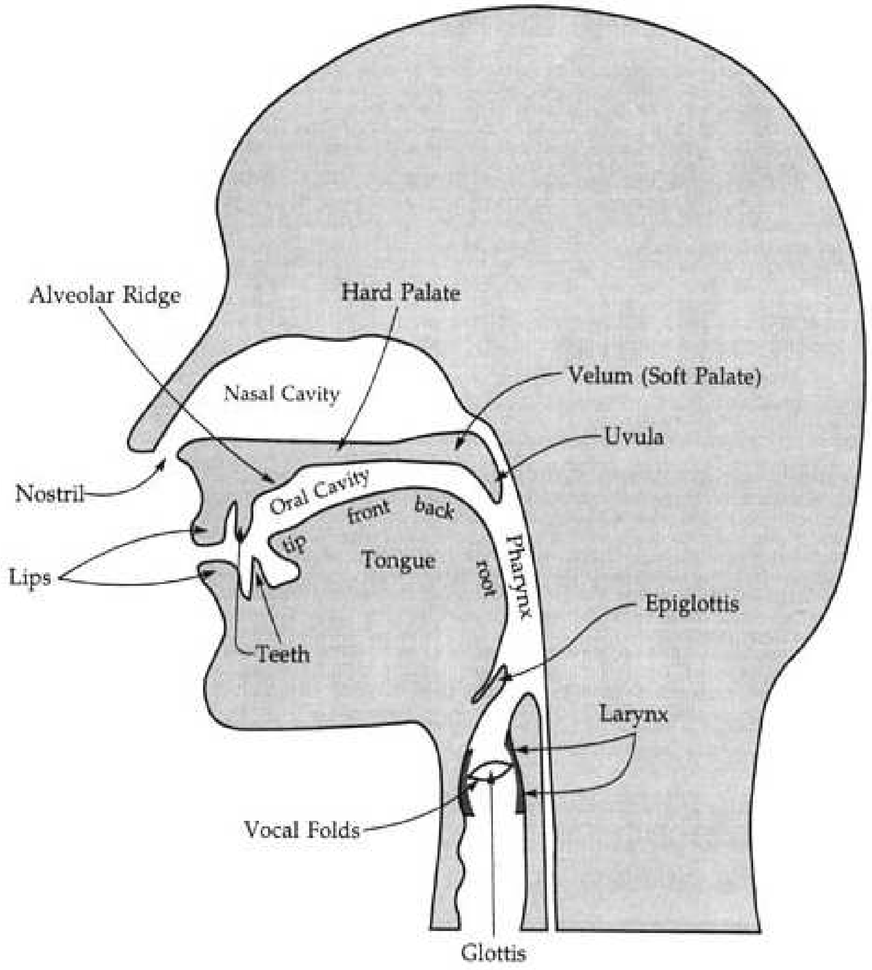
Voiced/Unvoiced Sounds

- Sounds made with vocal cords vibrating: **voiced**
  - E.g. /g/, /d/, etc.
  - All English vowel sounds are voiced
- Sounds made without vocal cord vibration: **voiceless**
  - E.g. /k/, /t/, etc.

Consonants (voicing/place/manner)

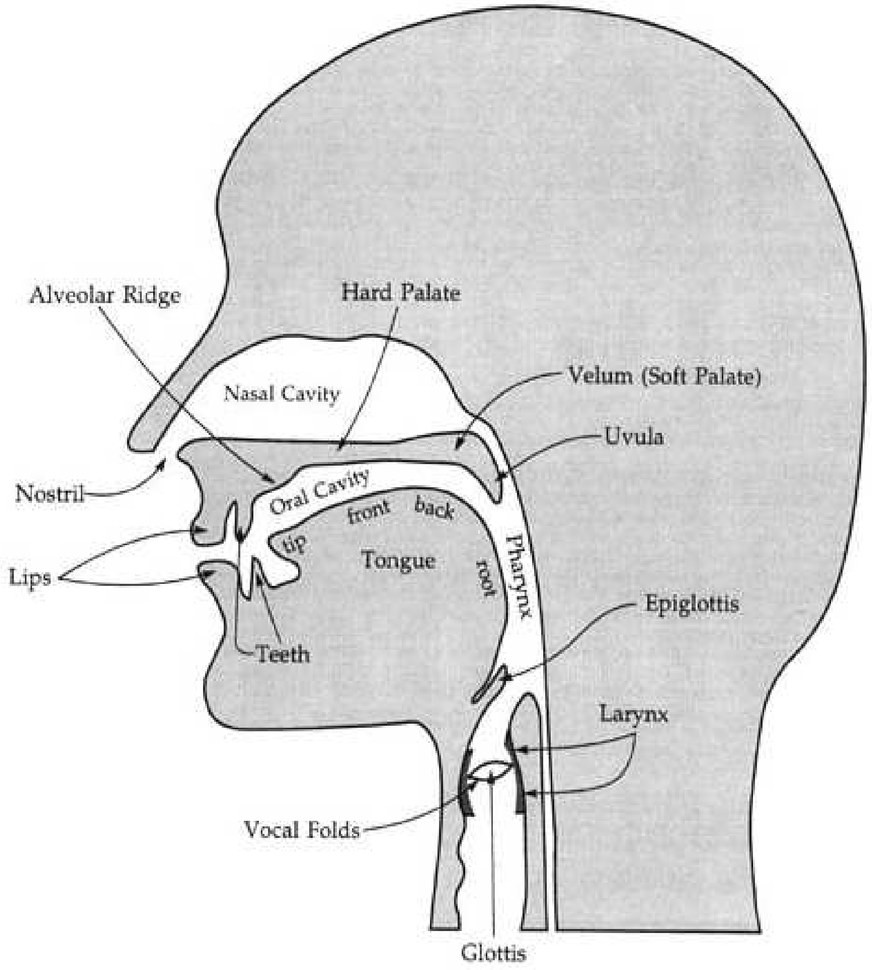
- “Consonants are made by restricting or blocking the airflow in some way, and may be voiced or unvoiced.” (J&M, Ch. 7)
- Consonants can be labeled depending on
  - *where* the constriction is made (place of articulation)
  - *how* the constriction is made (manner of articulation)

Place of articulation



- Bilabial (both lips)
  [b],[p],[m], etc.
- Labiodental (with lower lip and upper teeth)
  [f], [v], etc.
- Interdental (tip of tongue between teeth)
  [ɵ] (<u>th</u>ought), [ð] (<u>th</u>is)

Manner of articulation



- Plosive/Stop (airflow completely blocked followed by a release)
  [p],[g],[t],etc.

- Fricative (constricted airflow)
  [f], [s], [th], etc.

- Affricate (stop + fricative)
  [ch], [jh], etc.

- Nasal (lowering velum)
  [n], [m], etc.

See realtime MRI productions of vowels and consonants here: http://sail.usc.edu/span/rtmri_ipa/je_2015.html