CS 224S / Linguist 285
Spoken Language Processing

Andrew Maas | Stanford University | Spring 2024

Lecture 2: Phonetics

Announcements

- **Homework 1 Available on the website**
  - Due on Monday April 15 at 11:59pm Pacific
- **Homework is Colab and written section**
  - Today's lecture will help with phonetic transcription!
  - Phonetic transcription can be ambiguous
    - In Homework 1 we give points for multiple correct answers when there is ambiguity
    - Use only the restricted set of phonemes in Arpabet (not full IPA)
- **Office hours:**
  - Andrew's on Wednesdays after class on the patio outside (including today)
  - TA office hours start next week

Outline

- Phonetics Overview
- ARPAbet Phonetic Transcription
- Articulatory Phonetics: How we produce sounds
- Acoustic Phonetics: How we produce and visualize sound waves
- Overview of Prosody: Conveying meaning beyond just the words we say

Phonetics Overview

Phonetics Overview

- **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
- **Some vocabulary:**
  - Phone: Any distinct speech sound or gesture
  - Phoneme: A speech sound that conveys meaning (a syllable or word would change if the phoneme were swapped)
  - Allophone: A distinct speech sound that does not affect word meaning (i.e. variations of sounds within the same phoneme category)

Do we need phonetics to build systems that accurately process spoken language?

- **Modern systems (based on deep learning) are far less reliant on encoding phonetic domain knowledge directly than previous approaches**
  - Allowing deep learning models to learn letter-sound mappings from data can perform much better than hand engineering phonetic structure into a recognition or synthesis system
- **However …**
- **Basic understanding of phonetics and speech production helps with describing and debugging spoken language systems**
  - E.g. how does an accent change the sound of pronunciations?
- **Phonetic categories are not arbitrary. They model the biology of *how* humans produce speech**
  - Understanding the space of possible speech sounds gives a nice perspective on comparing spoken languages across the world, and how they evolve

ARPAbet Transcription

- An alphabet for transcribing American English phonetic sounds
- Prominent because a lot of early speech recognition research focused on English
- ARPAbet does not contain many sounds that occur in languages other than English

English Vowels

In ARPAbet

	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”

https://corpus.linguistics.berkeley.edu/acip/

Articulatory Parameters for English Consonants

In ARPAbet

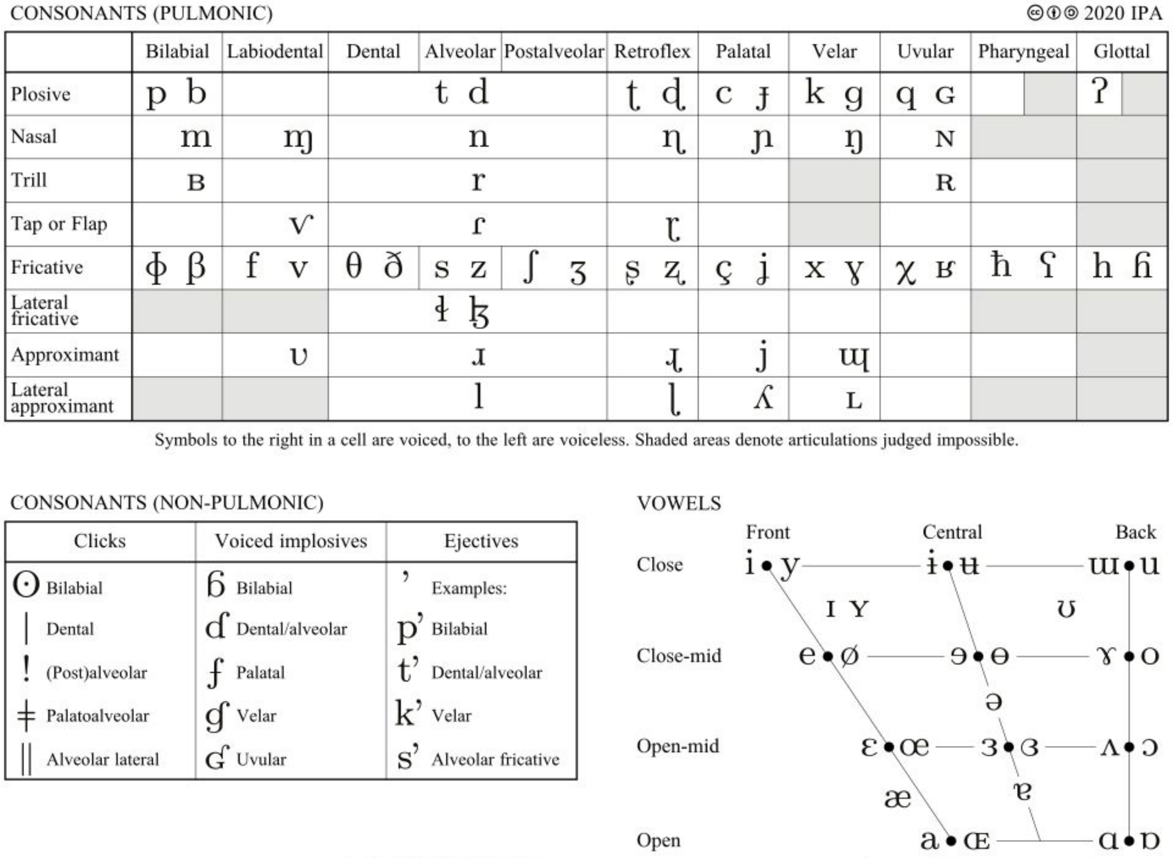
Manner of articulation	Place of articulation													
	bilabial		labiodental		inter-dental		alveolar		palatal		velar		glottal	
stop	p	b					t	d			k	g	q	
fric.			f	v	th	dh	s	z	sh	zh			h	
affric.									ch	jh				
nasal		m						n				ng		
approx		w						l/r		y				
flap							dx							

□ Voiceless
▨ Voiced

Table 1: Jennifer Venditti

International Phonetic Alphabet (IPA)

Wikipedia IPA (with sounds)



CONSONANTS (PULMONIC)

2020 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

VOWELS

	Front	Central	Back
Close	i • y	ɨ • ʉ	ɯ • u
	ɪ ʏ		ʊ
Close-mid	e • ø	ɘ • ɵ	ɤ • o
		ə	
Open-mid	ɛ • œ	ɜ • ɞ	ʌ • ɔ
	æ	ɐ	
Open	a • ɶ		ɑ • ɒ

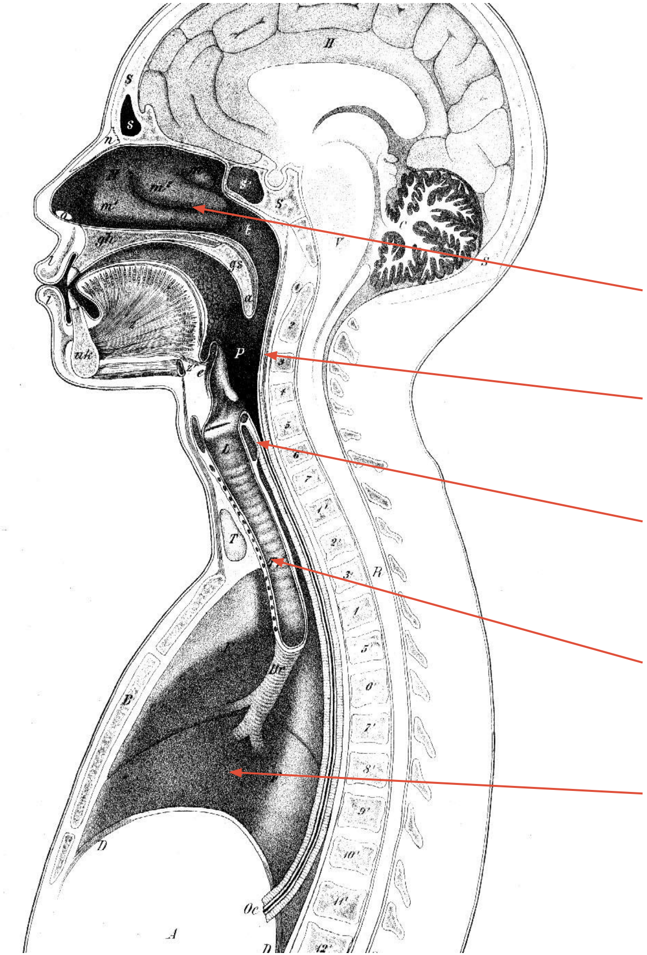
Articulatory Phonetics

- How speech sounds are made by articulators (moving organs)

Speech Production

- **Flow:** we (normally) speak while breathing out. Respiration provides airflow. "Pulmonic egressive airstream"
  - Airstream sets vocal folds in motion. Vibration of vocal folds produces sounds. Sound is then modulated by:
- **Resonance:** shape of vocal tract causing harmonics
- **Articulation:** manipulation of airflow
  - Oral tract: uvula, soft palate (velum), hard palate, tongue, lips, teeth
  - Nasal tract

Sagittal section of the vocal tract



Nasal Cavity

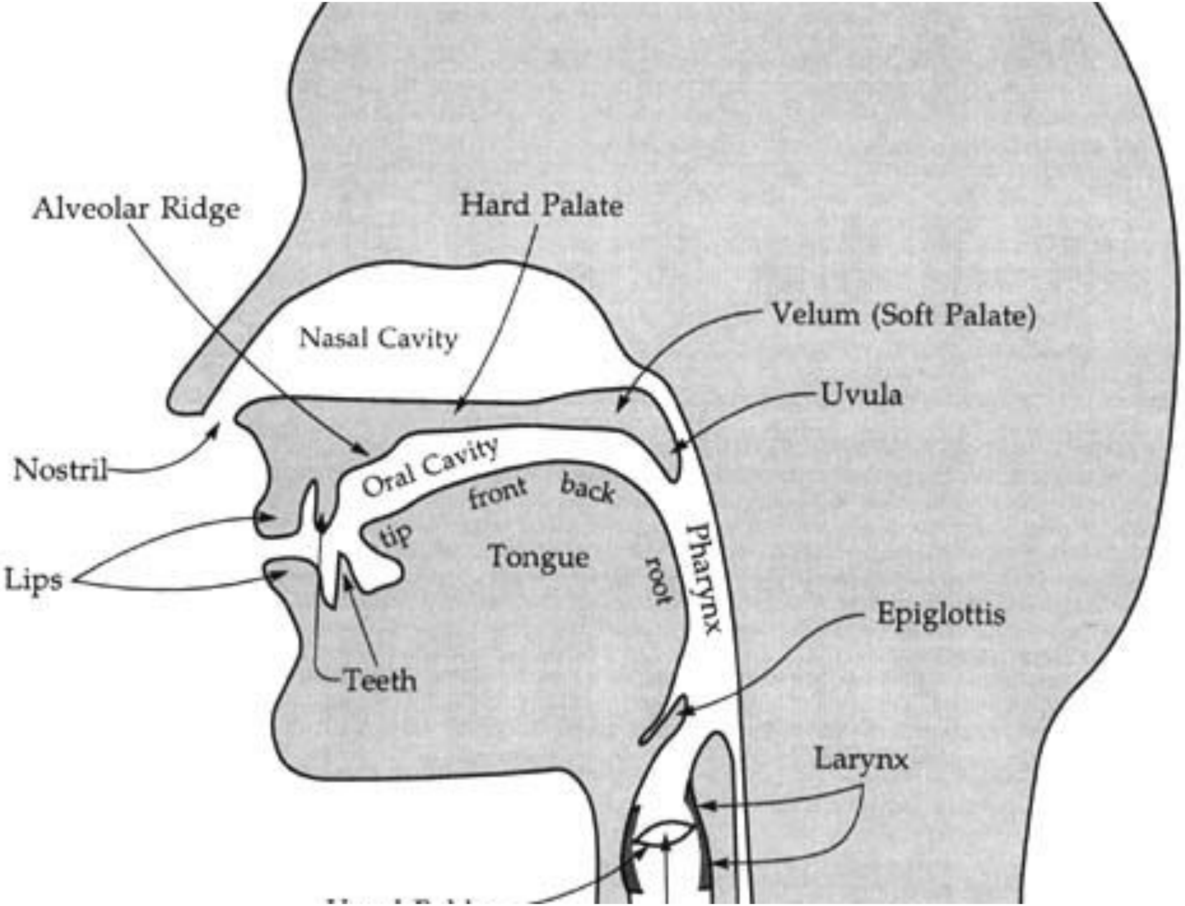
Pharynx

Vocal Folds
(within the Larynx)

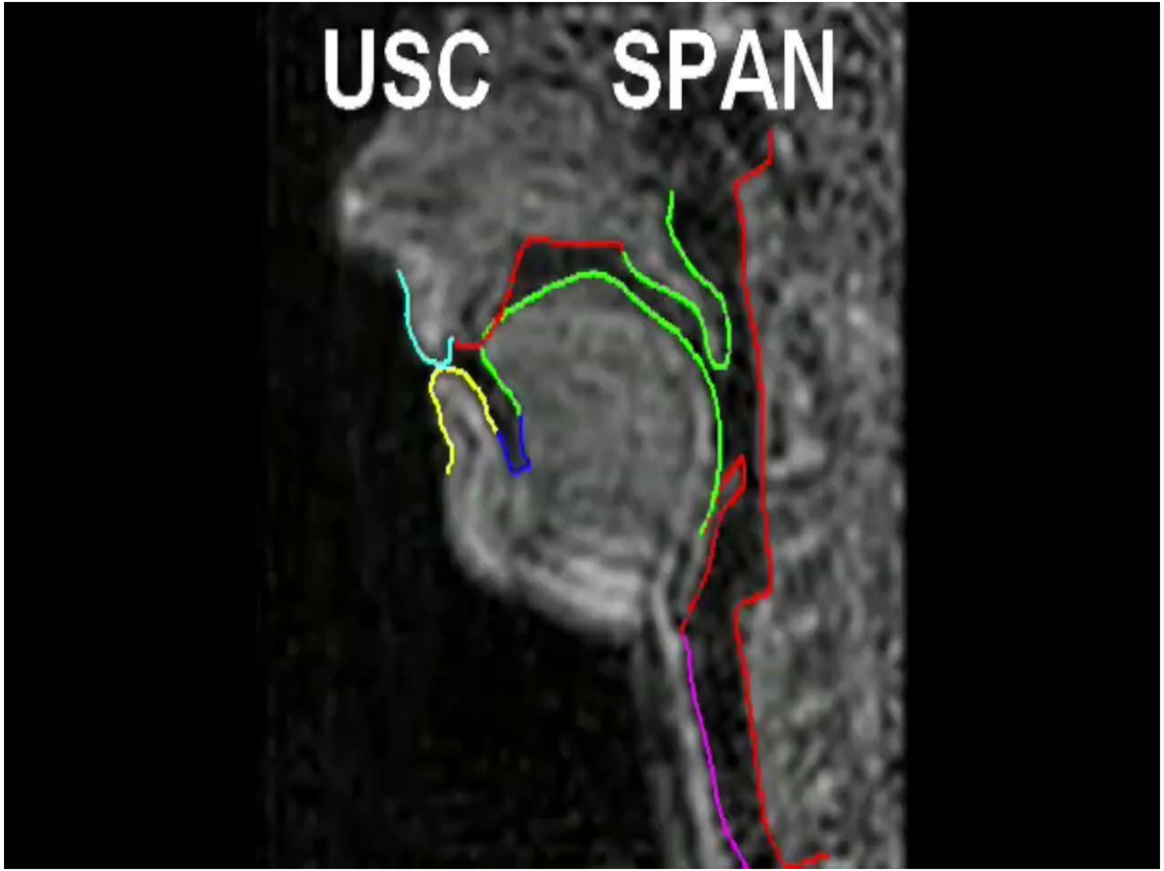
Trachea

Lungs

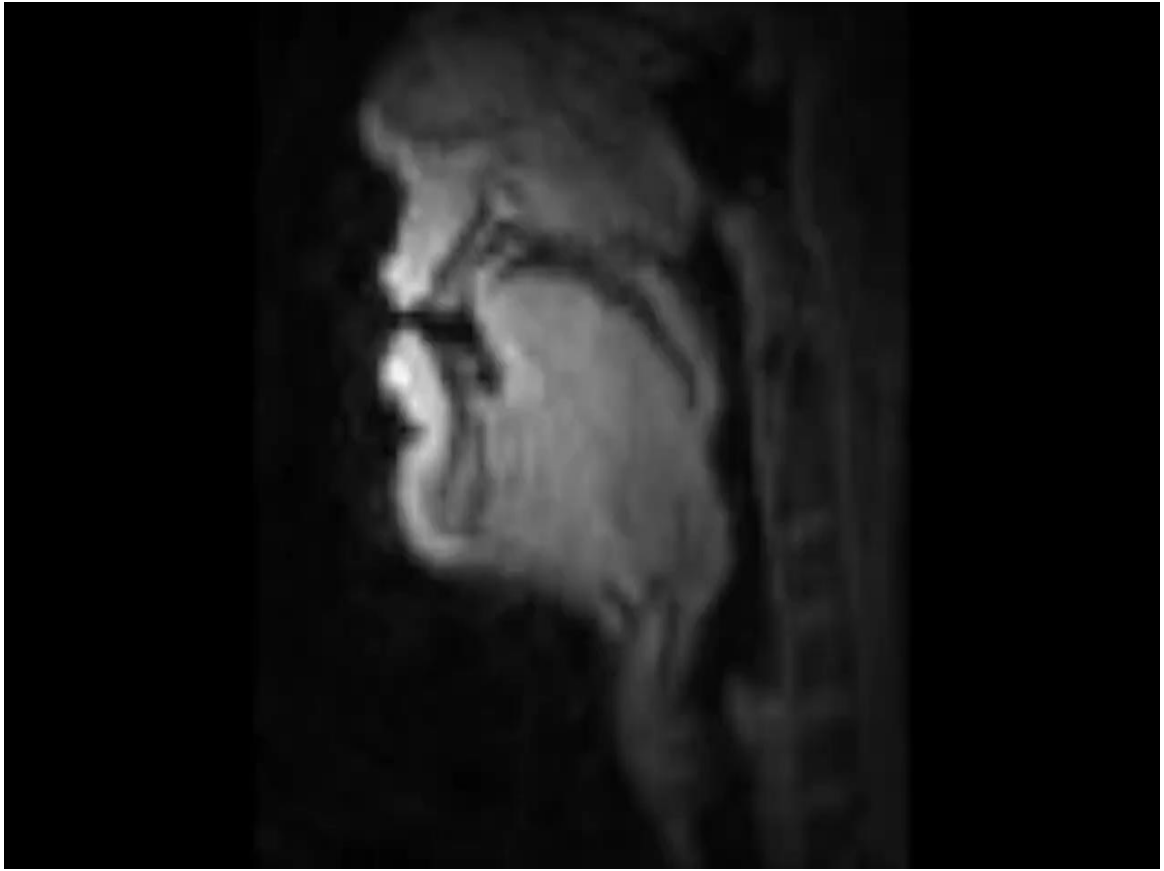
Sagittal section of the vocal tract



USC's SAIL Lab
Shri Narayanan

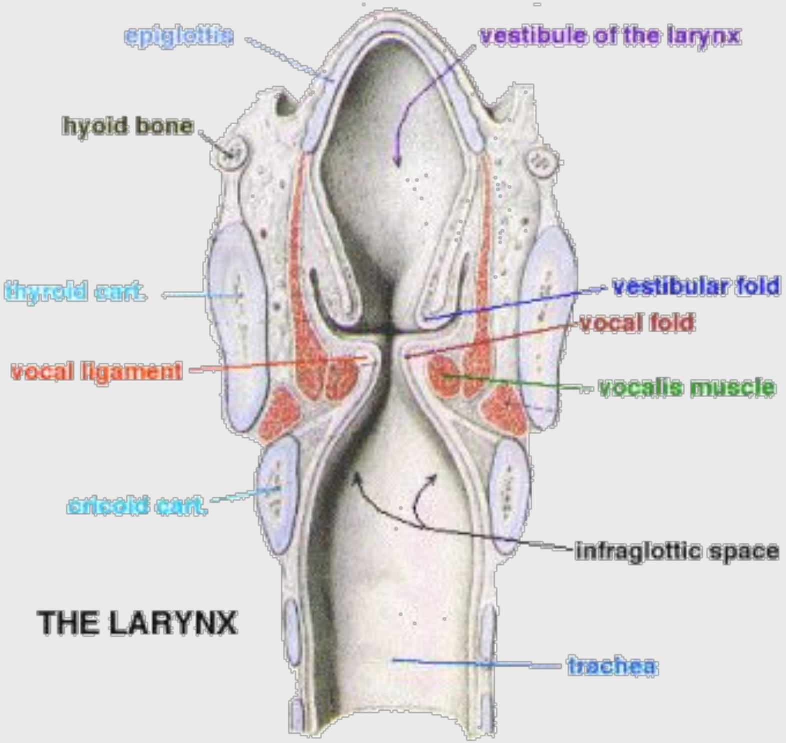


Tamil



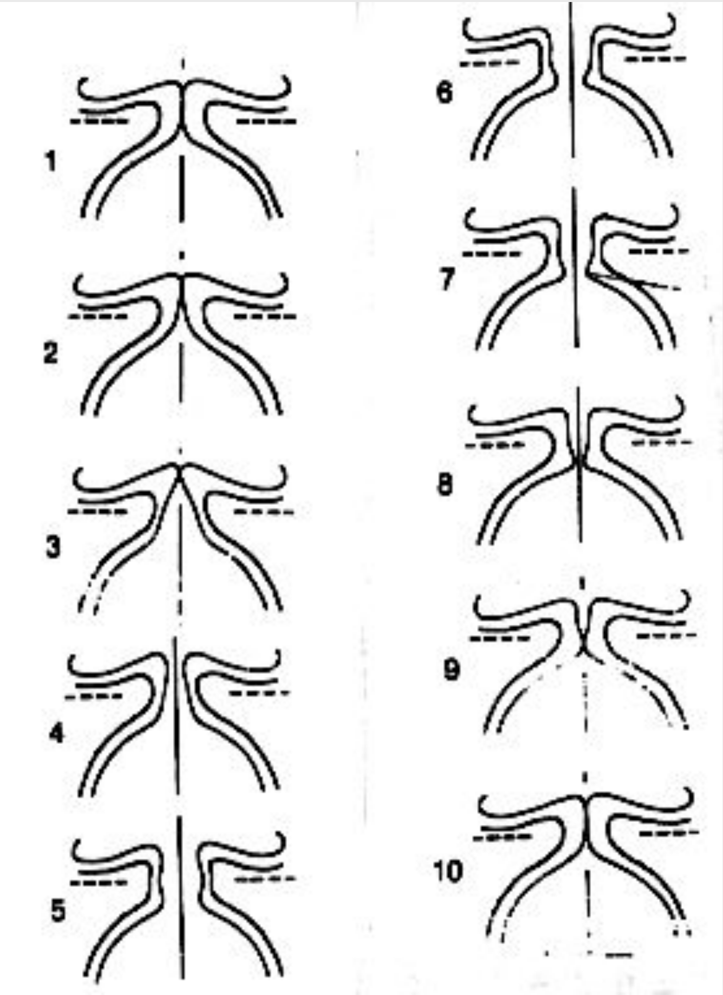
Larynx and Vocal Folds

- 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)



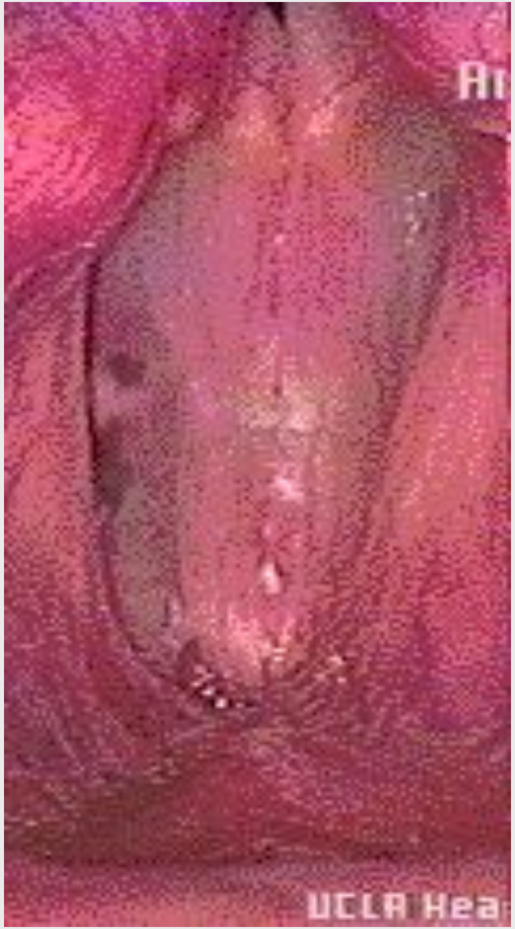
Voicing

- 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



Vocal Fold Vibration

- 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



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

Consonants and Vowels

Consonants:

phonetically, sounds with audible noise produced by a constriction

Vowels:

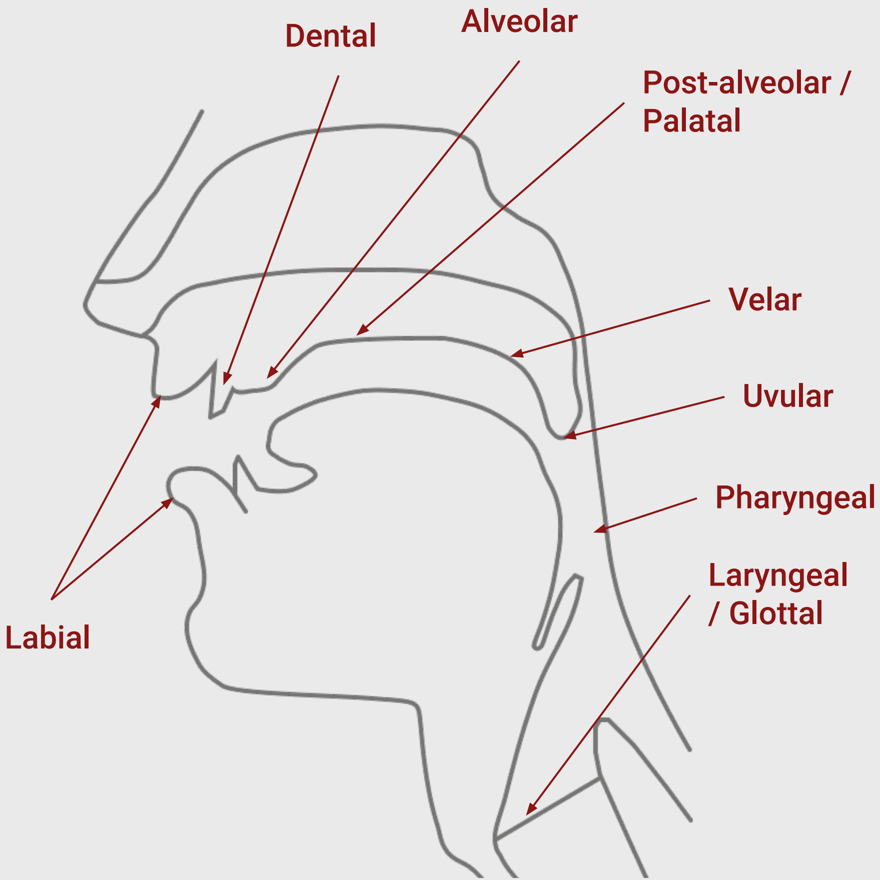
phonetically, sounds with no audible noise produced by a constriction

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

USC: Soprano Singing

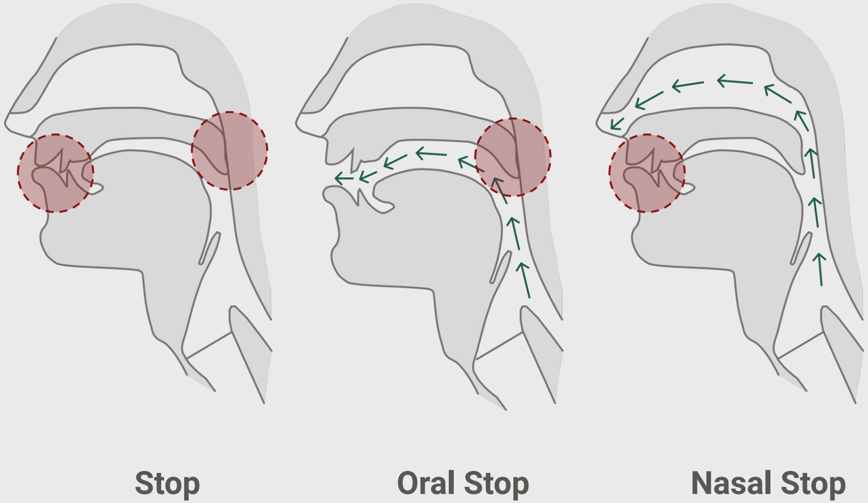
Place of Articulation

- 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)



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

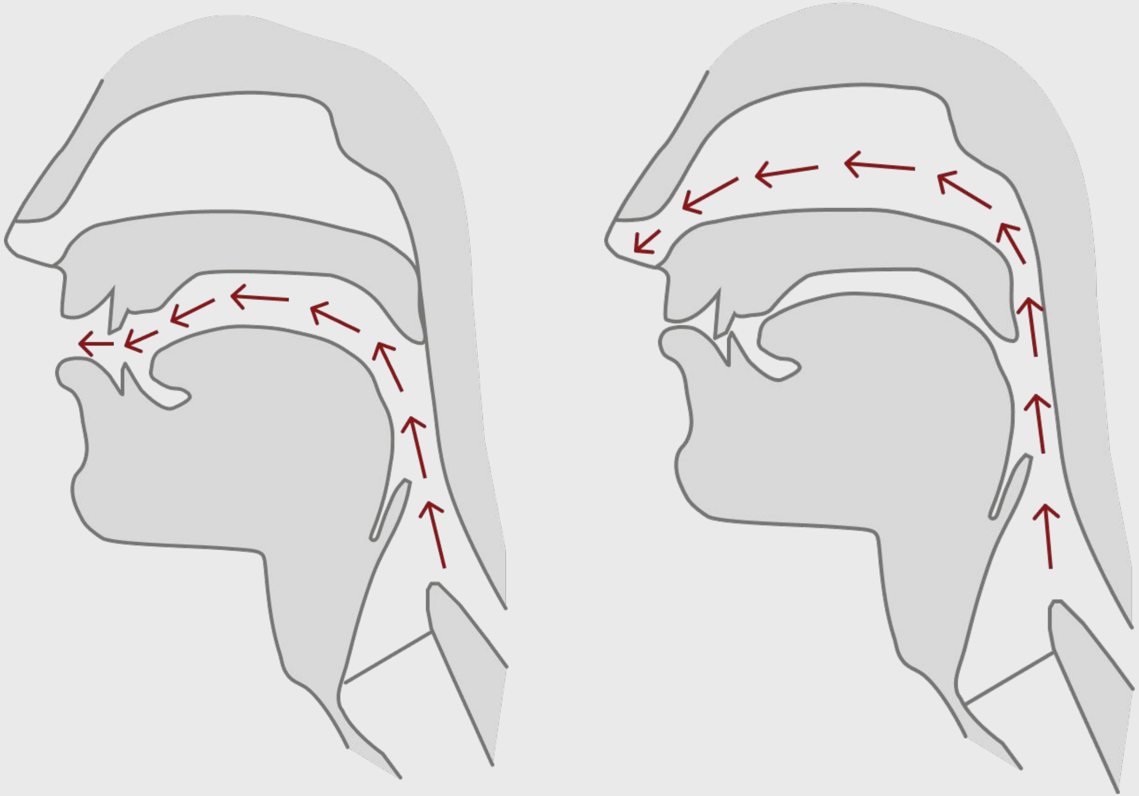


Stop

Oral Stop

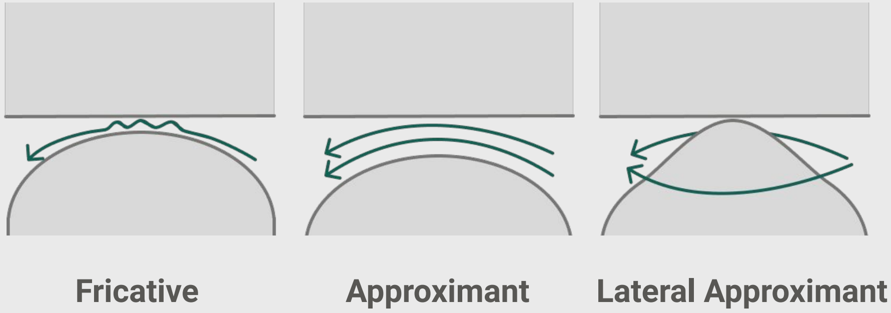
Nasal Stop

Oral vs Nasal Sounds

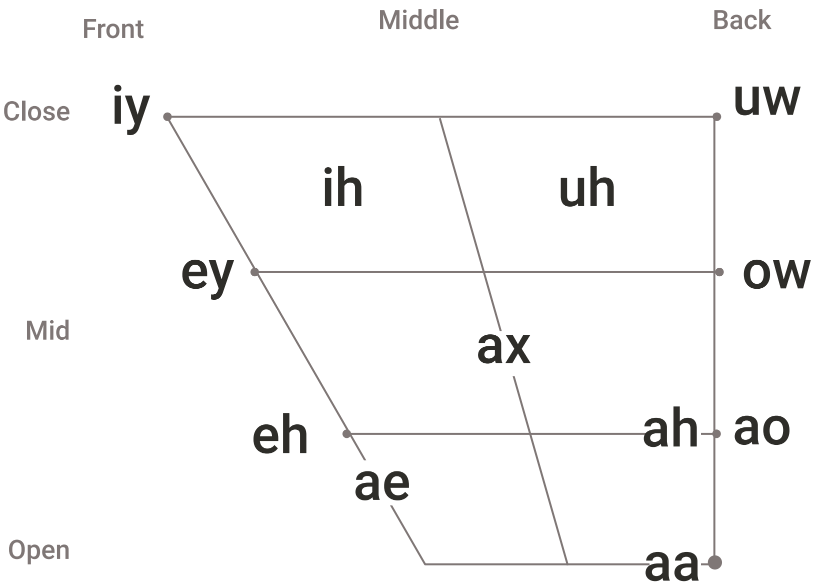


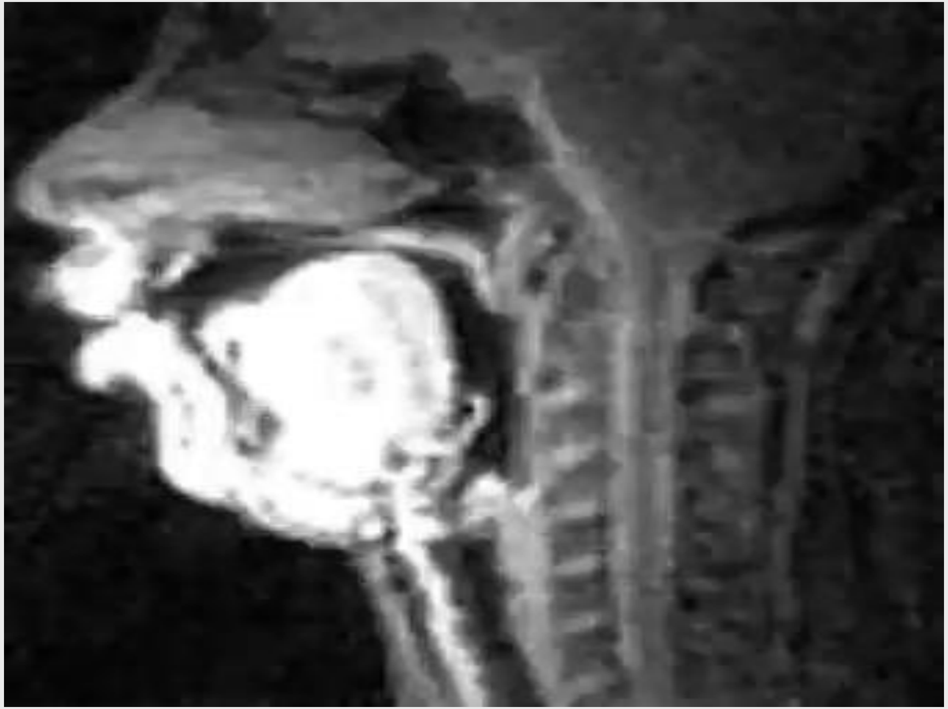
More on Manner of Articulation of Consonants

- **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



Tongue Position for Vowels





Articulatory Parameters for English Consonants

In ARPAbet

	Place of articulation													
Manner of articulation	bilabial		labiodental		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							

□ Voiceless
▨ Voiced

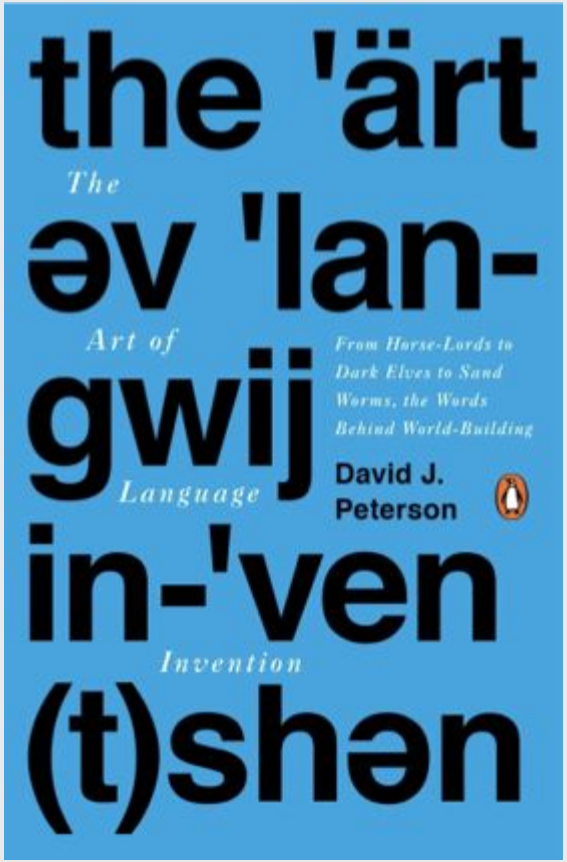
Table 1: Jennifer Venditti

The Art of Language Invention

- Fun, informative book on phonetics and phonotactics across languages.

- Great audio book!

- Talk Video


the 'ärt
The
əv 'lan-
Art of
gwij
Language
in-'ven
Invention
(t)shən
From Horse-Lords to Dark Elves to Sand Worms, the Words Behind World-Building
David J. Peterson

Acoustic Phonetics

- Acoustic properties of speech sounds

Sound Waves are Longitudinal Waves

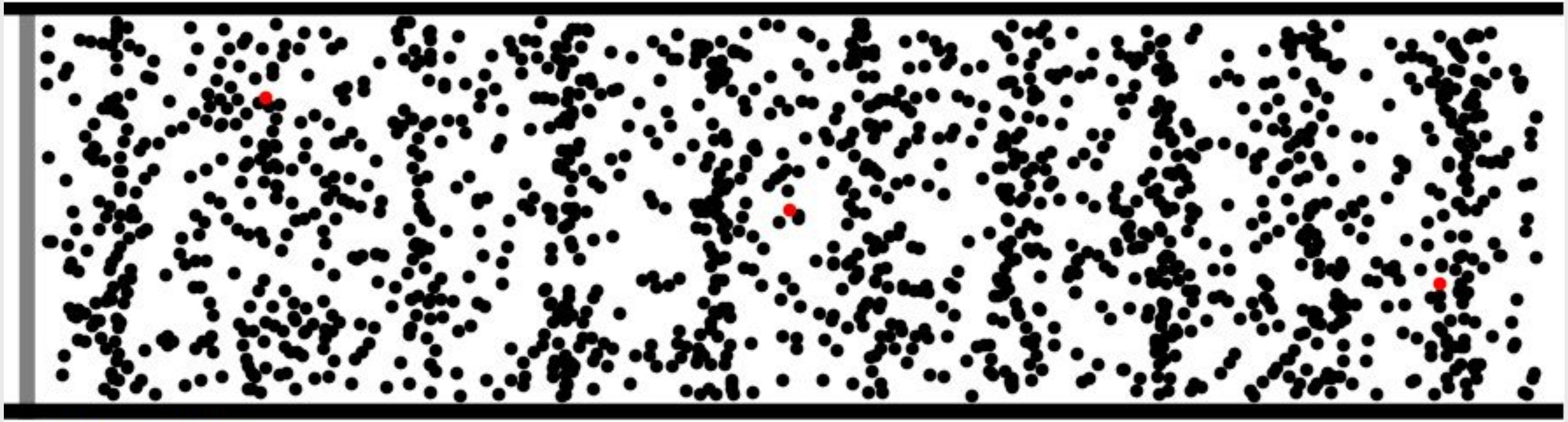
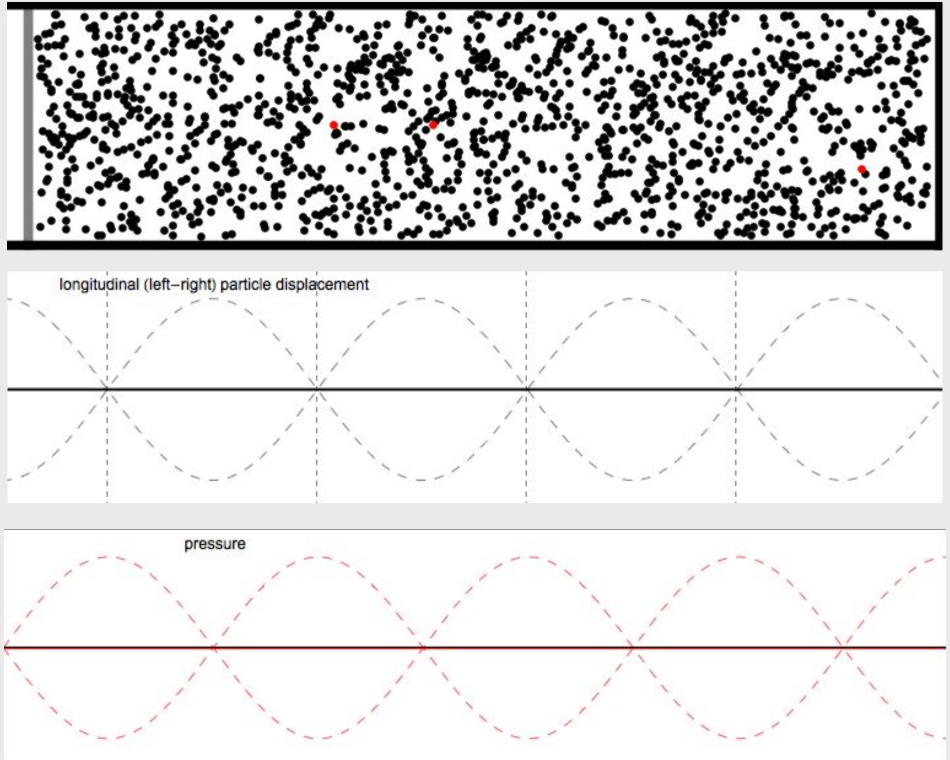


Image: Dan Russell (2011)

Sound Waves are Longitudinal Waves



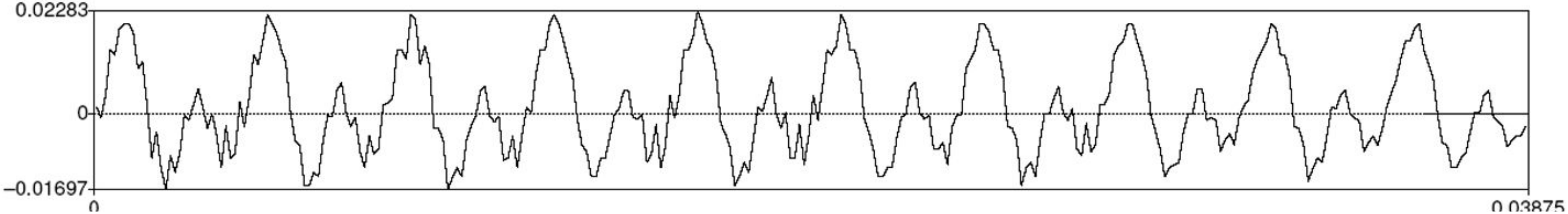
Particle Displacement

Pressure

Image: Dan Russell (2011)

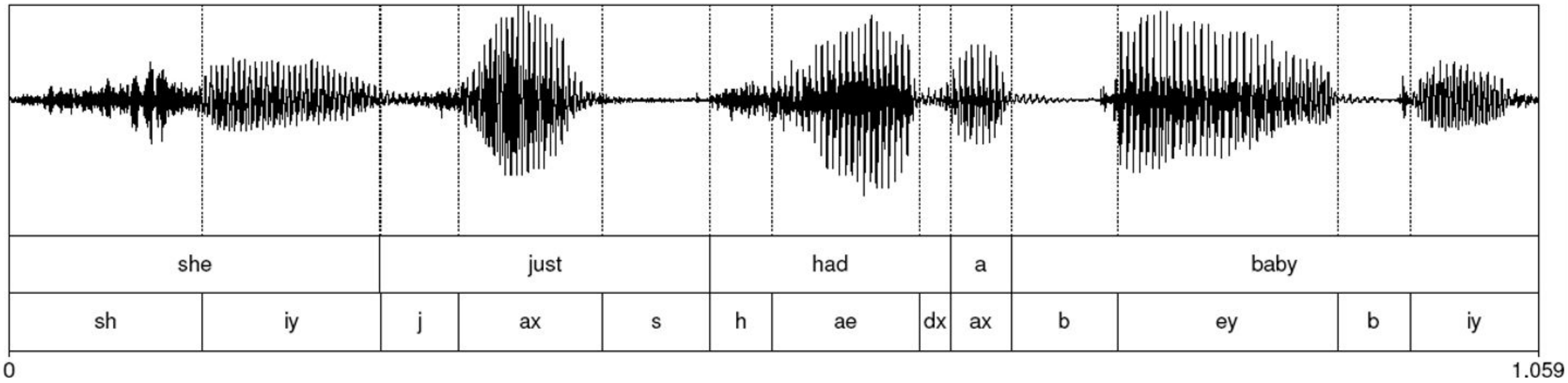
Back to Waves: Fundamental Frequency

- Waveform of the vowel [iy]



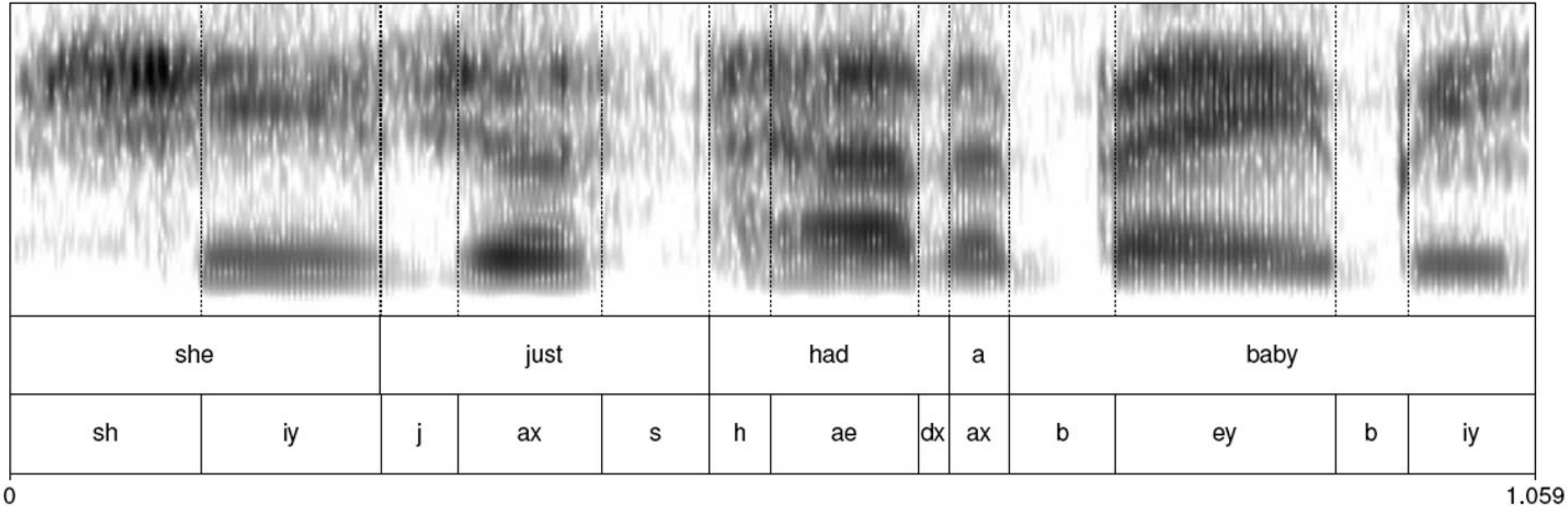
- Frequency: 10 repetitions / .03875 seconds = 258 Hz
- This is speed that vocal folds move, hence voicing
- Each peak corresponds to an opening of the vocal folds
- The low frequency of the complex wave is called the fundamental frequency of the wave or F0

She Just Had a Baby



- Note that vowels all have regular amplitude peaks
- Stop consonant
- Closure followed by release
- Notice the silence followed by slight bursts of emphasis: very clear for [b] of “baby”
- Fricative: noisy. [sh] of “she” at beginning

Spectrogram: Spectrum + Time Dimension



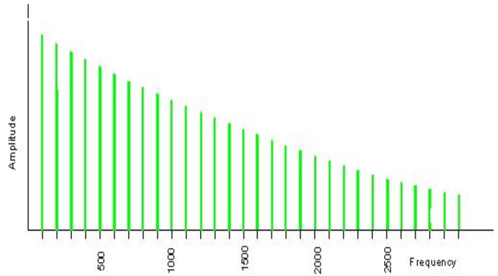
Source Filter Model of Vowels

- Any body of air will vibrate in a way that depends on its size and shape
- Vocal tract as "amplifier"; amplifies certain harmonics
- Formants are result of different shapes of vocal tract

Source Filter Model of Vowels

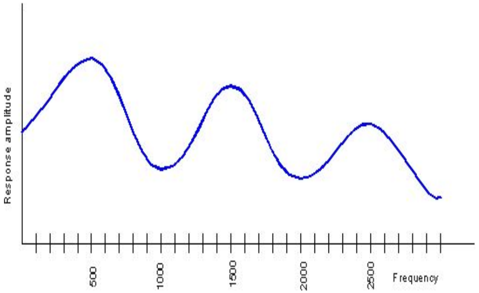
- **Source and filter are independent, so:**
  - Different vowels can have same pitch
  - The same vowel can have different pitch

Input



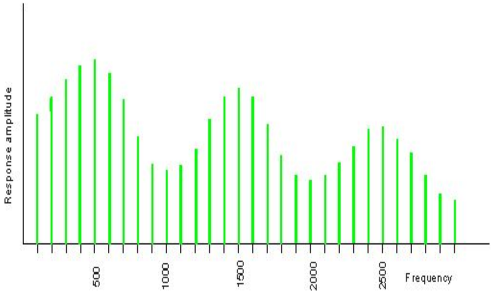
Glottal Spectrum

Filter



Vocal Tract Frequency Response Function

Output



Figures: Ratree Wayland

Resonances of the Vocal Tract

The human vocal tract as an open tube

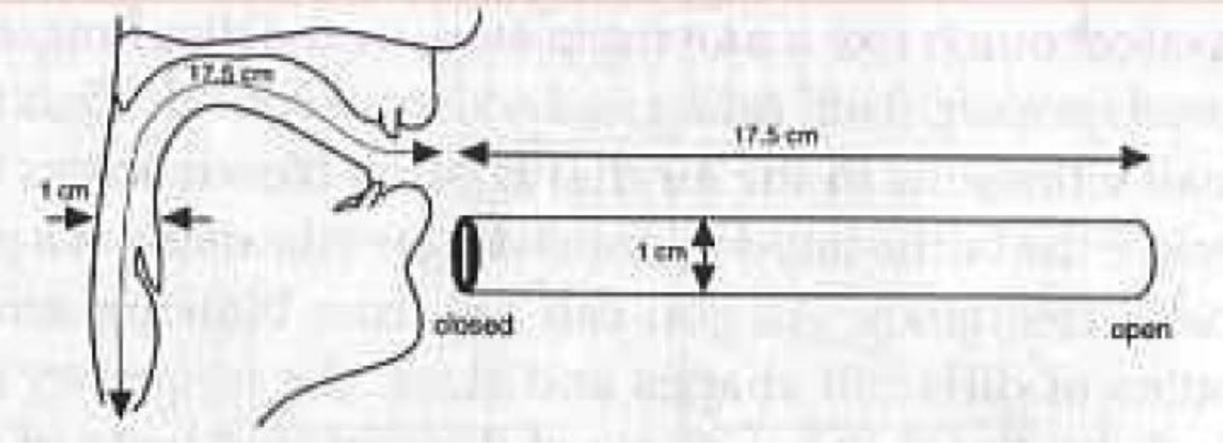
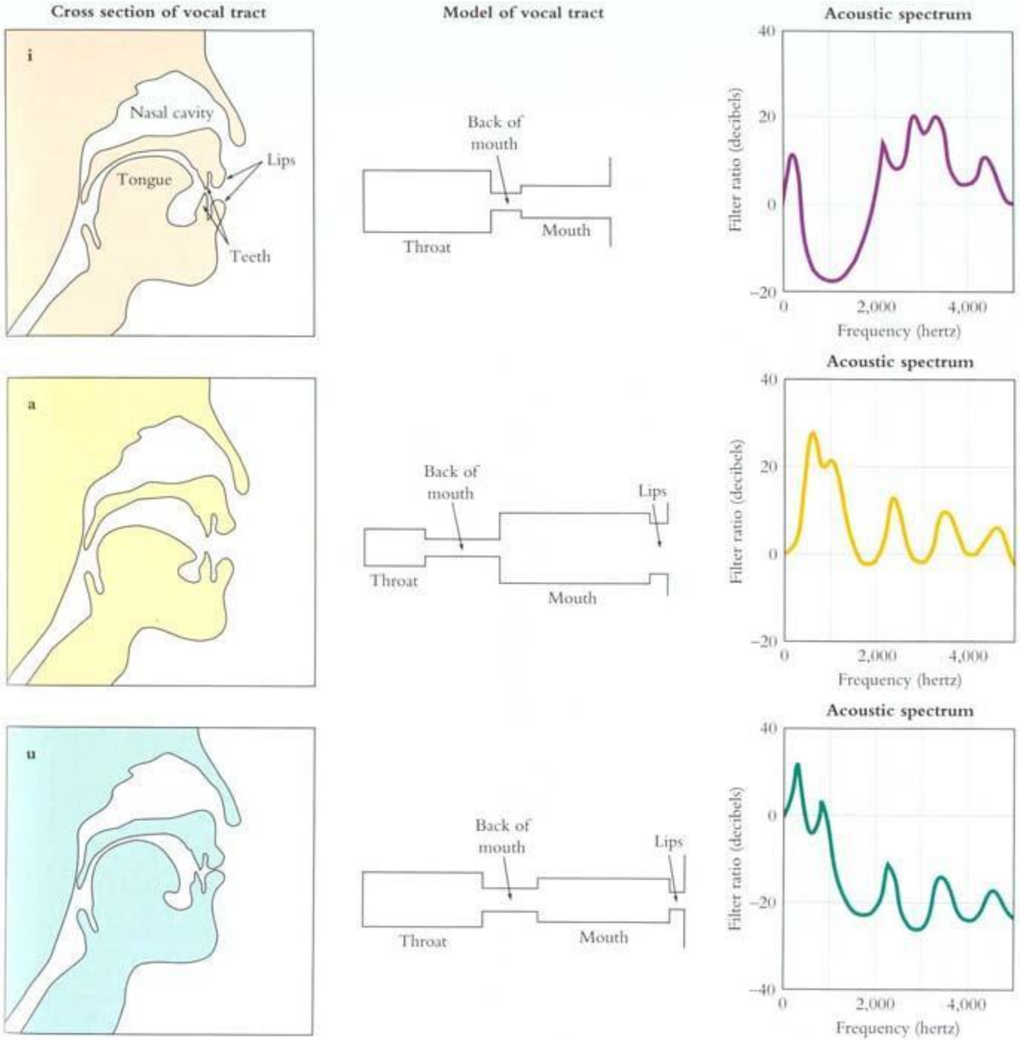


Figure: Ladefoged (1996) p.117

Resonances of the Vocal Tract

Figure: Mark Liberman

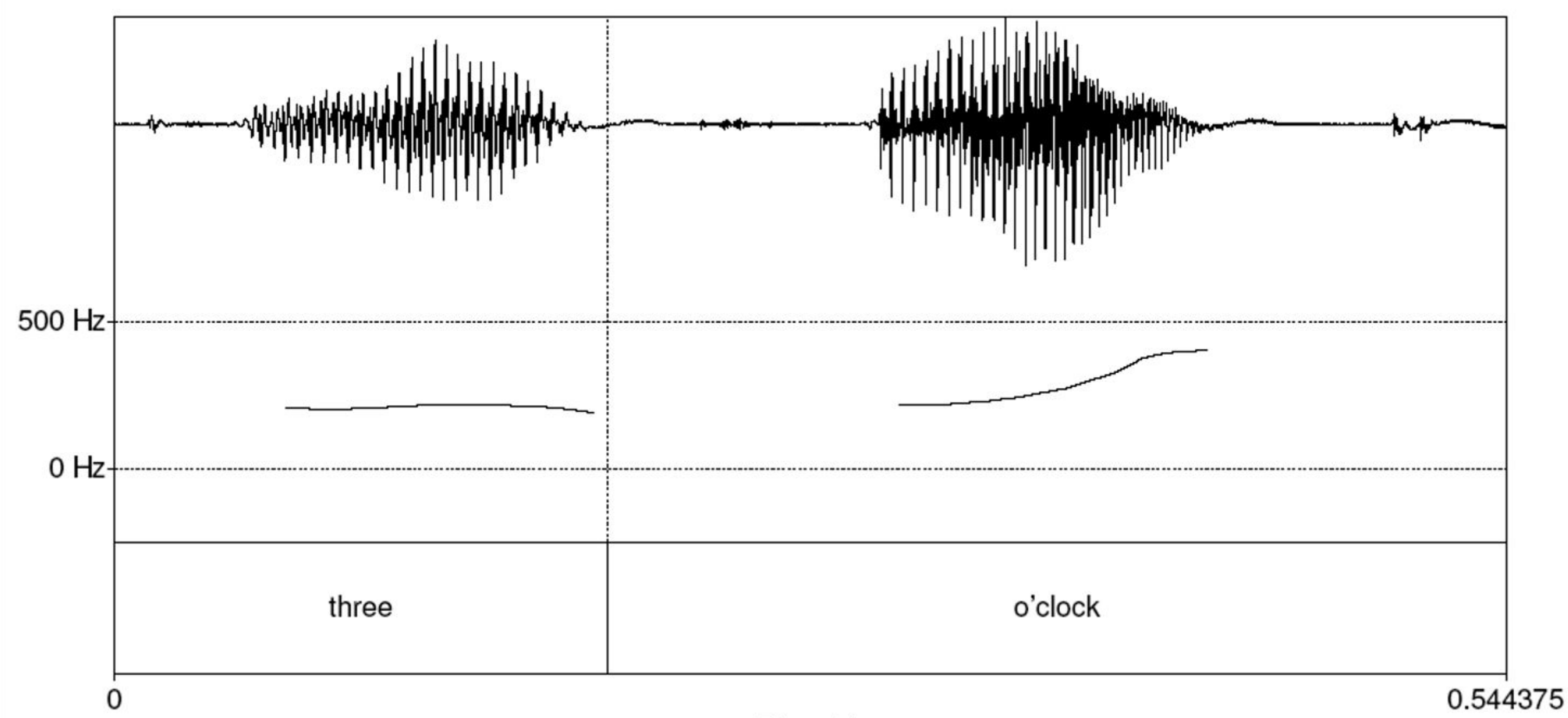


Prosody Overview

Defining Intonation

- **Ladd (1996) “Intonational phonology”**
- **“The use of suprasegmental phonetic features [...]**
  - Suprasegmental = above & beyond the segment/phone
    - F0 (pitch)
    - Intensity (energy)
    - Duration
- **to convey sentence-level pragmatic meanings”**
  - i.e. meanings that apply to phrases or utterances as a whole, not lexical stress, not lexical tone.

Pitch Track



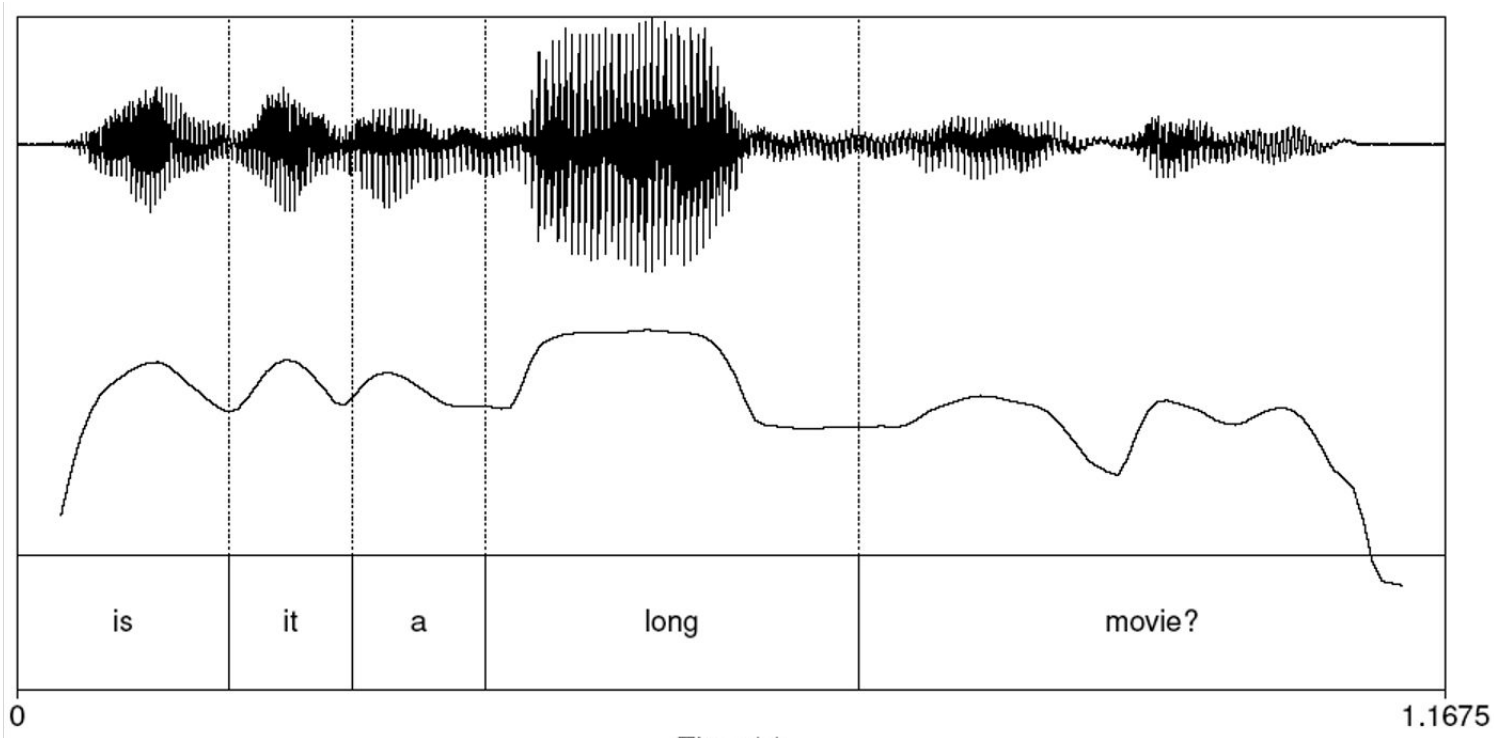
Pitch is not Frequency

- **Pitch is the mental sensation or perceptual correlate of F0**

- **Relationship between pitch and F0 is not linear;**
  - human pitch perception is most accurate between 100Hz and 1000Hz.
    - Linear in this range
    - Logarithmic above 1000Hz

- **Mel scale is one model of this F0-pitch mapping**
  - A mel is a unit of pitch defined so that pairs of sounds which are perceptually equidistant in pitch are separated by an equal number of mels
  - Frequency in mels = 1127 ln (1 + f/700)

Plot of Intensity



Three Aspects of Prosody

- **Prominence:** some syllables/words are more prominent than others
- **Structure/boundaries:** sentences have prosodic structure
  - Some words group naturally together
  - Others have a noticeable break or disjuncture between them
- **Tune:** the intonational melody of an utterance.

Prosodic Boundaries

I met Mary and Elena’s mother at the mall yesterday.

🔊

I met Mary, and Elena’s mother at the mall yesterday.

🔊

French [bread and cheese]

🔊

[French bread] and [cheese]

🔊

Thank You

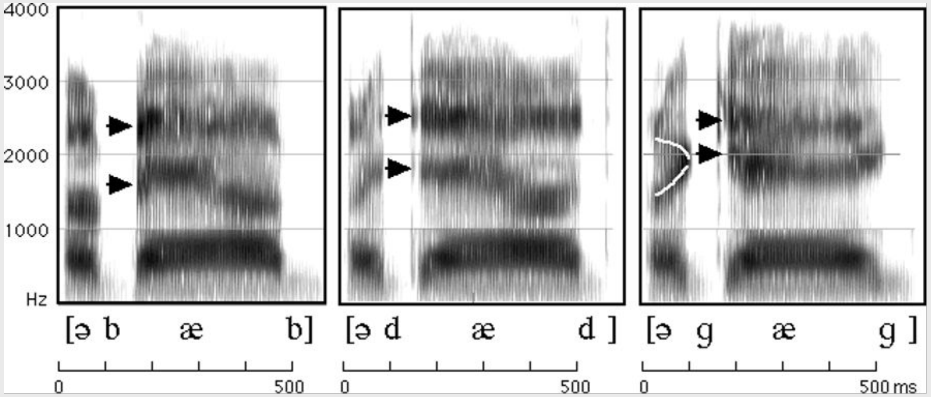
Appendix

Useful Links

- The ARPAbet
  - http://www.stanford.edu/class/cs224s/arpabet.html
- The CMU Pronouncing Dictionary
  - http://www.speech.cs.cmu.edu/cgi-bin/cmudict
- International Phonetic Alphabet:
  - http://en.wikipedia.org/wiki/International_Phonetic_Alphabet

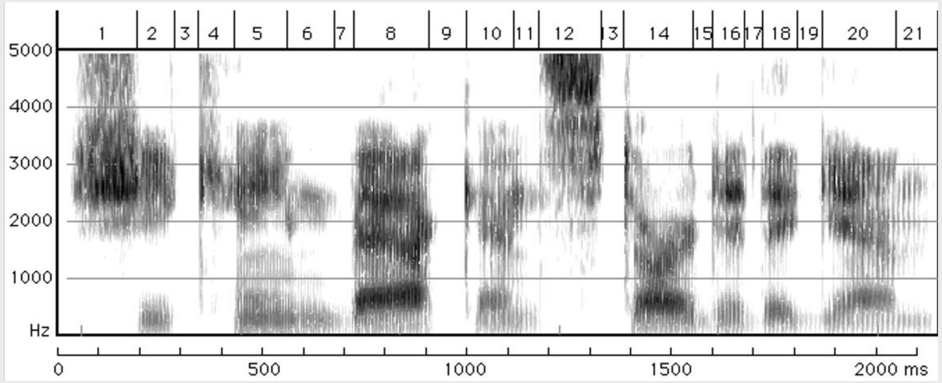
How to Read Spectrograms

- **bab:** closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab”
- **dad:** first formant increases, but F2 and F3 slight fall
- **gag:** F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials

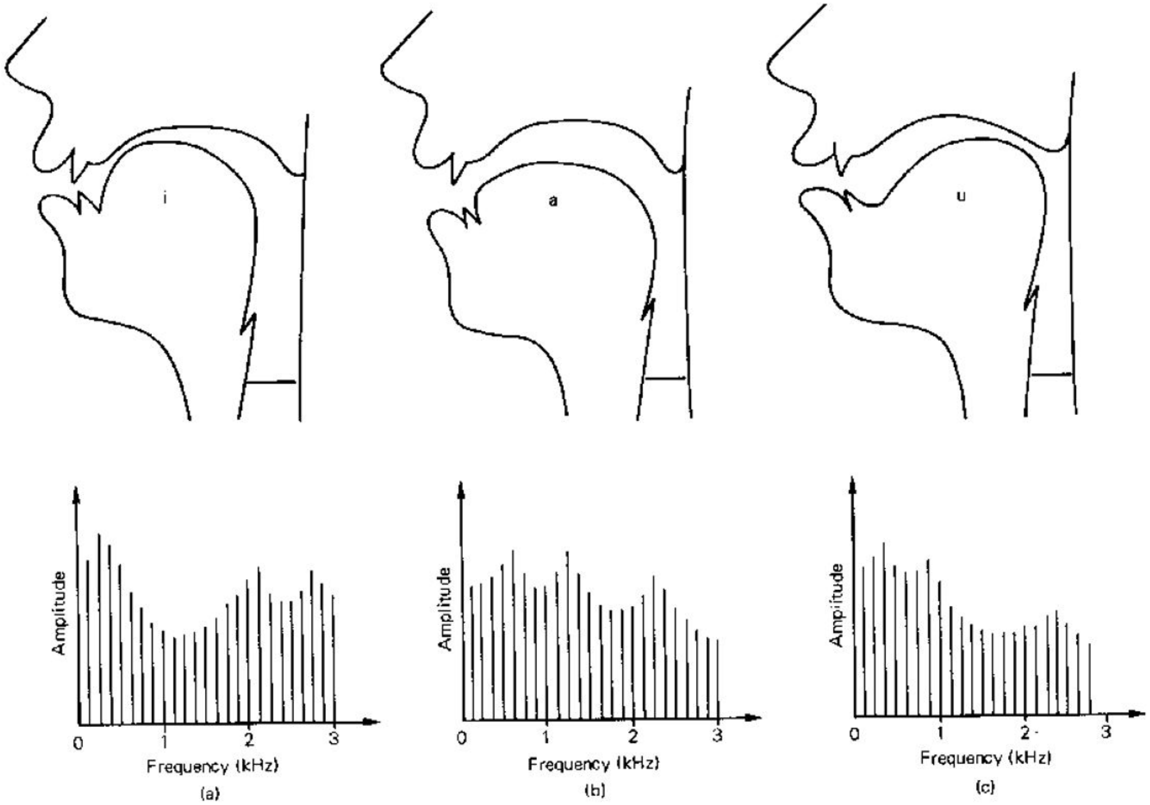


She Came Back and Started Again

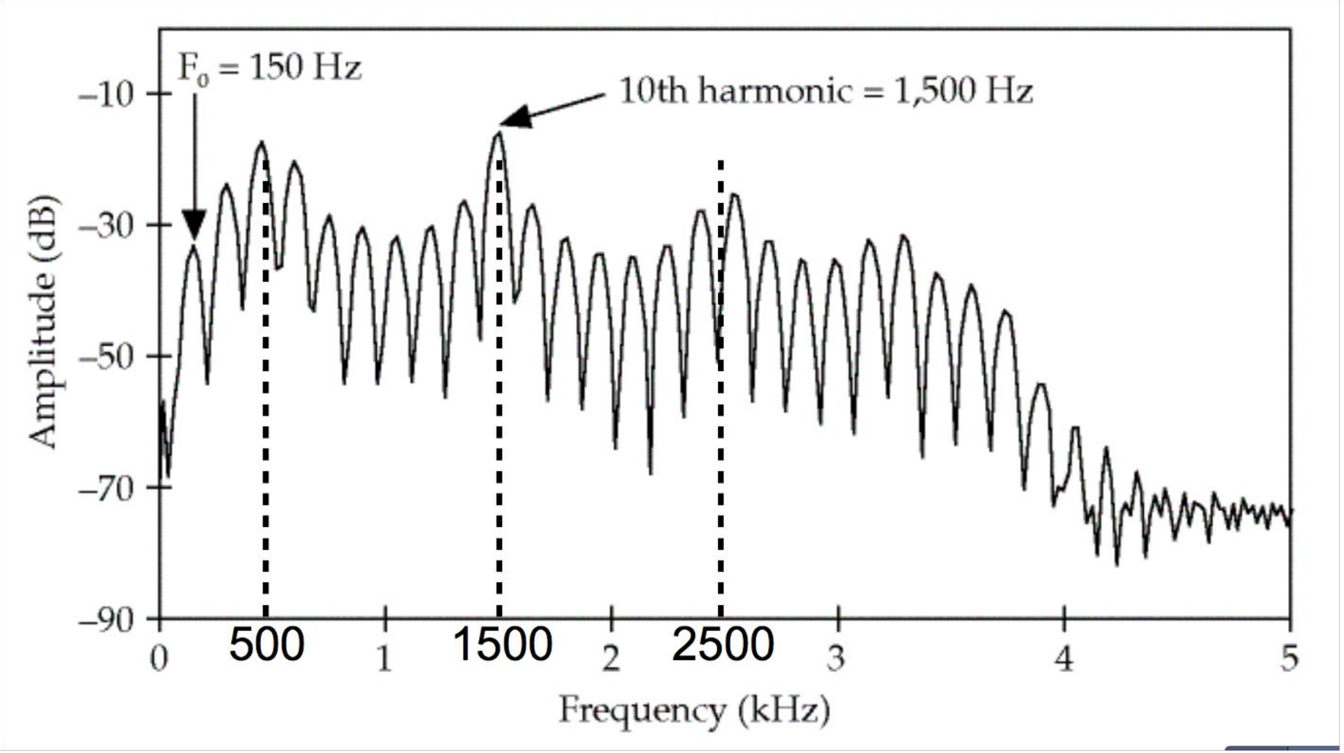
- Lots of high-freq energy
- Closure for k
- Burst of aspiration for k
- [ey] faint 1100 Hz formant is nasalization
- Bilabial nasal
- Short b closure, voicing barely visible.
- [ae] note upward transitions after bilabial stop at beginning
- Note F2 and F3 coming together for "k



Vowels

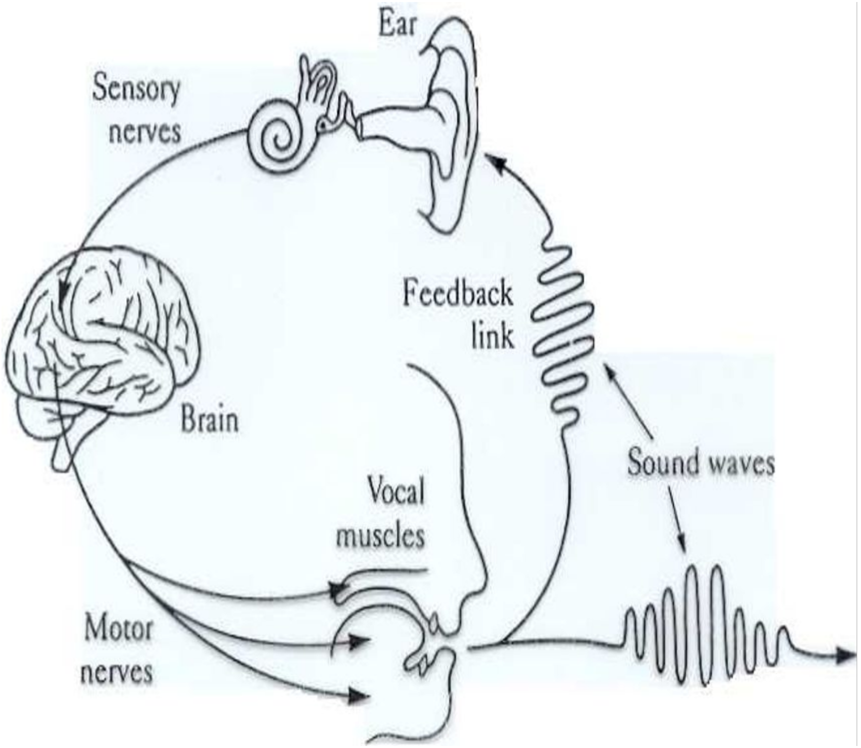

i
a
u
Amplitude
0
1
2
3
Frequency (kHz)
(a)
Amplitude
0
1
2
3
Frequency (kHz)
(b)
Amplitude
0
1
2
3
Frequency (kHz)
(c)

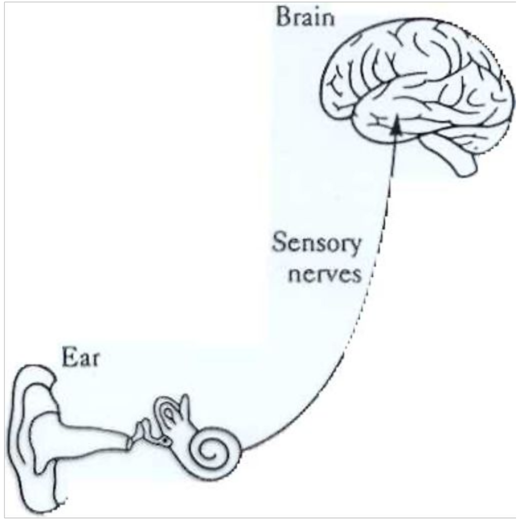
The Oral Cavity Amplifies Some Harmonics



The Speech Chain (Denes and Pinson)

Speaker





Hearer

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