CS 224S / LINGUIST 285
Spoken Language Processing

Andrew Maas
Stanford University
Spring 2017

Lecture 2: Phonetics

Original slides by Dan Jurafsky

Homework 1

- Out after lecture today. Due in 1 week
- PDF handout linked on website syllabus
- You'll need to download PRAAT; details are in the homework.

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

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

- The CMU Pronouncing Dictionary
- http://www.speech.cs.cmu.edu/cgi-bin/cmudict

- What about other languages?
- International Phonetic Alphabet:
- http://en.wikipedia.org/wiki/International_Phonetic_Alphabet

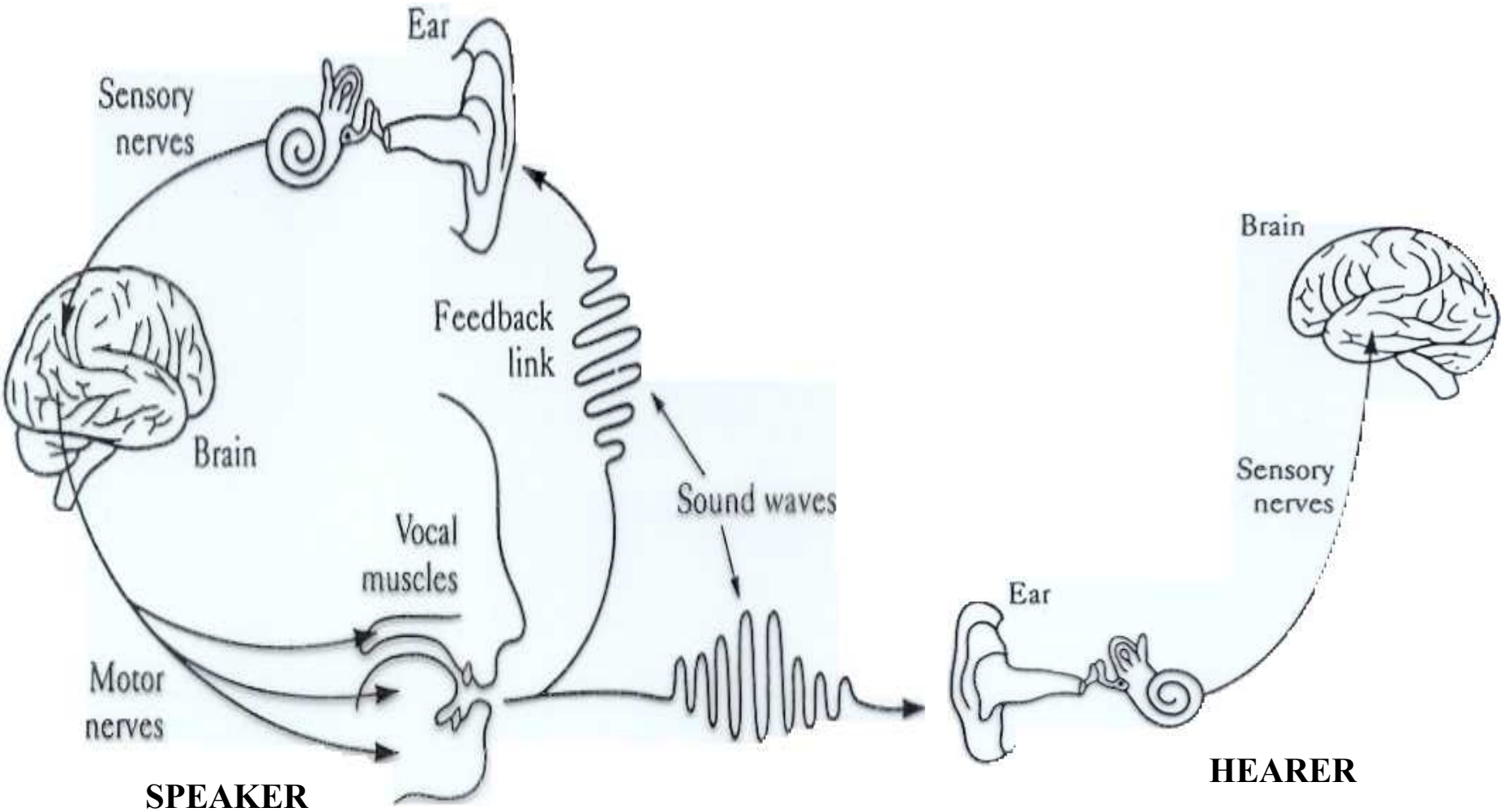
ARPAbet Vowels

	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"

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

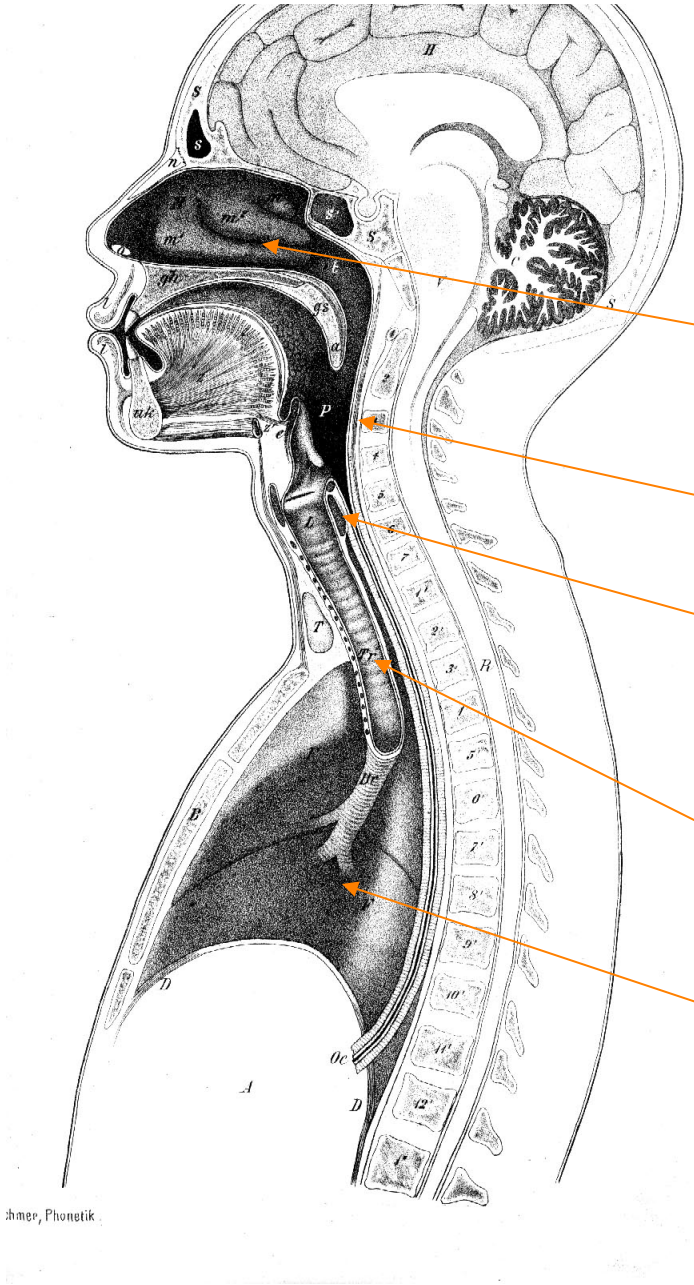
The Speech Chain (Denes and Pinson)



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)

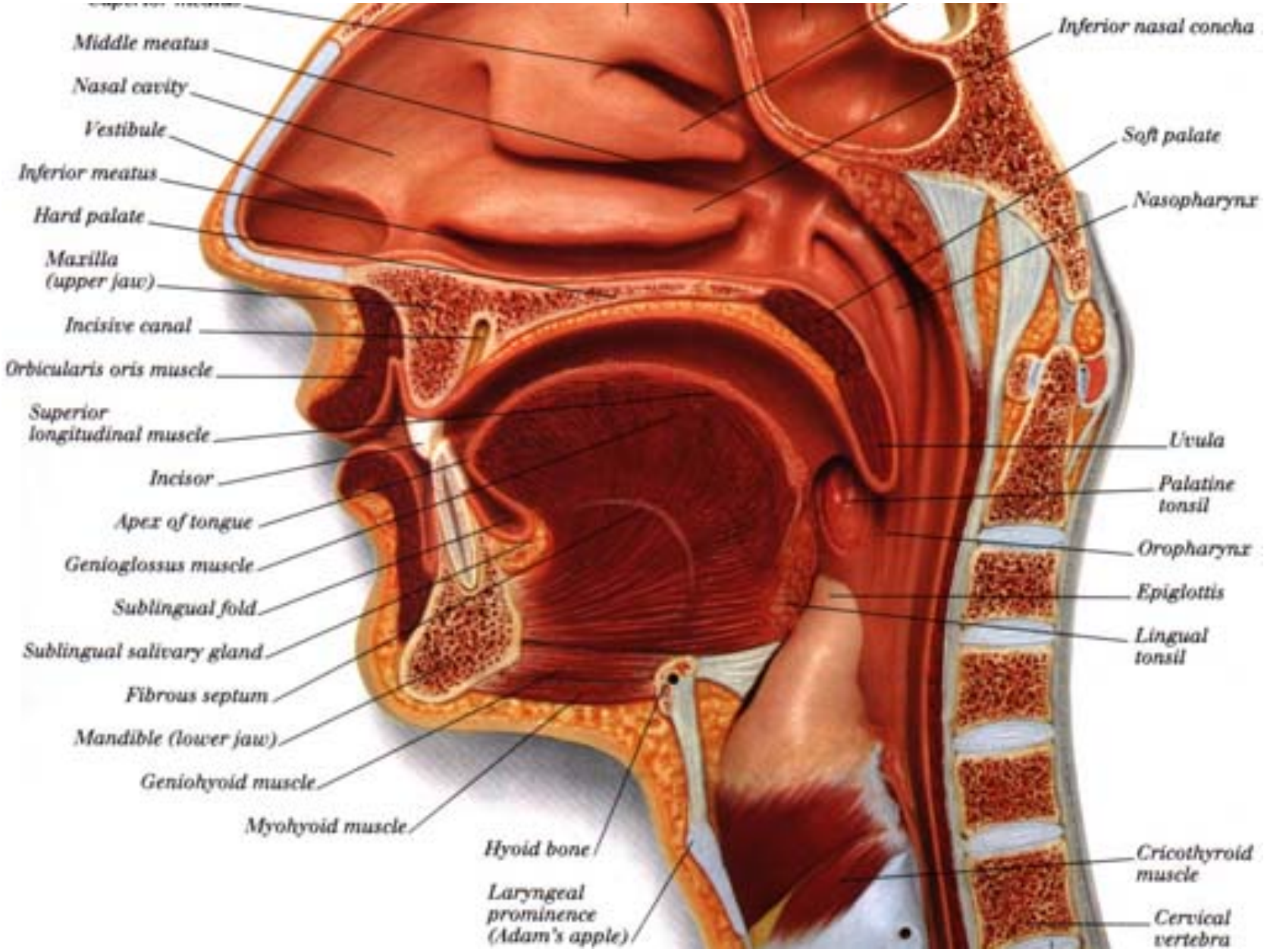
Nasal Cavity

Pharynx

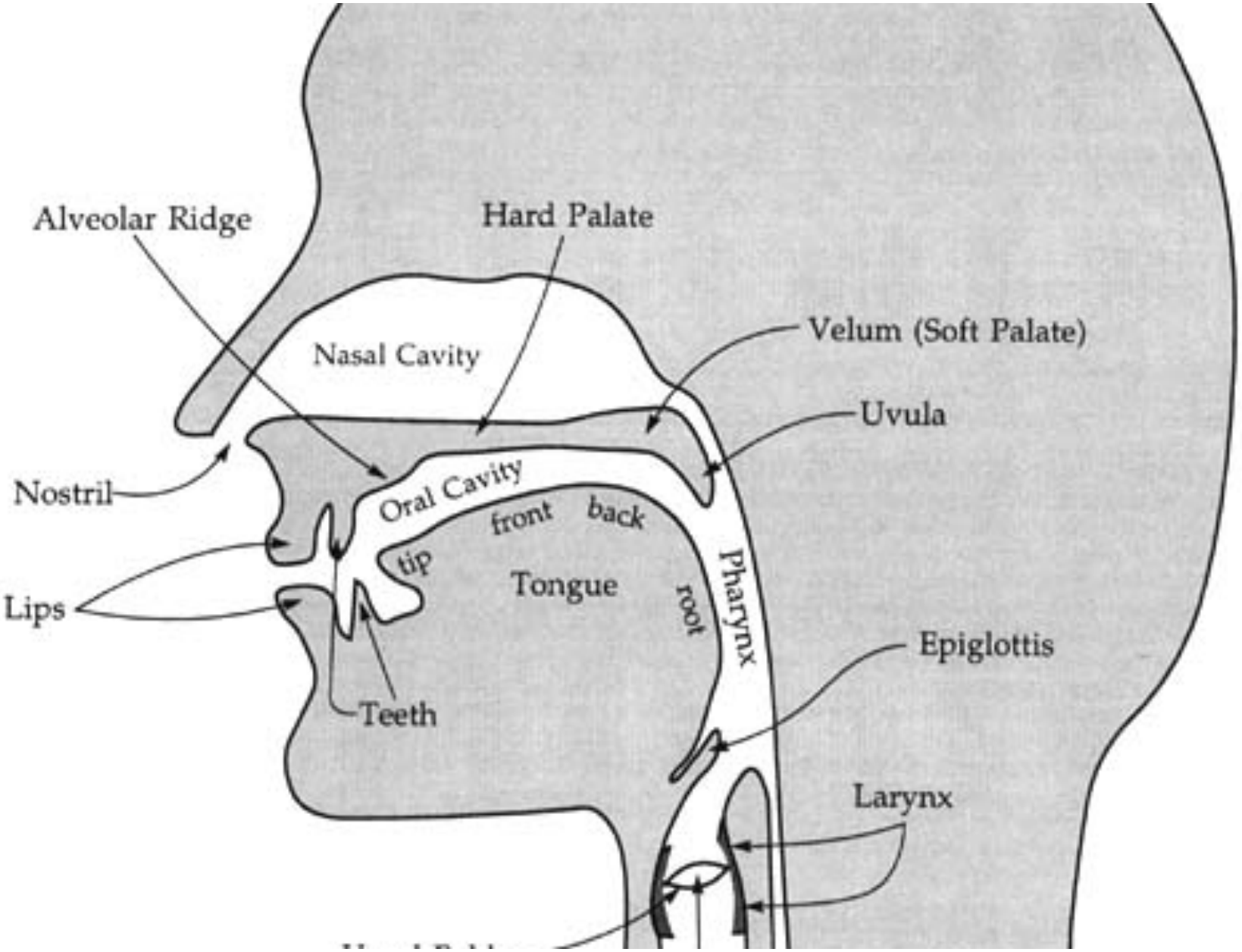
Vocal Folds (within the Larynx)

Trachea

Lungs



From Mark Liberman's website, from Ultimate Visual Dictionary



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

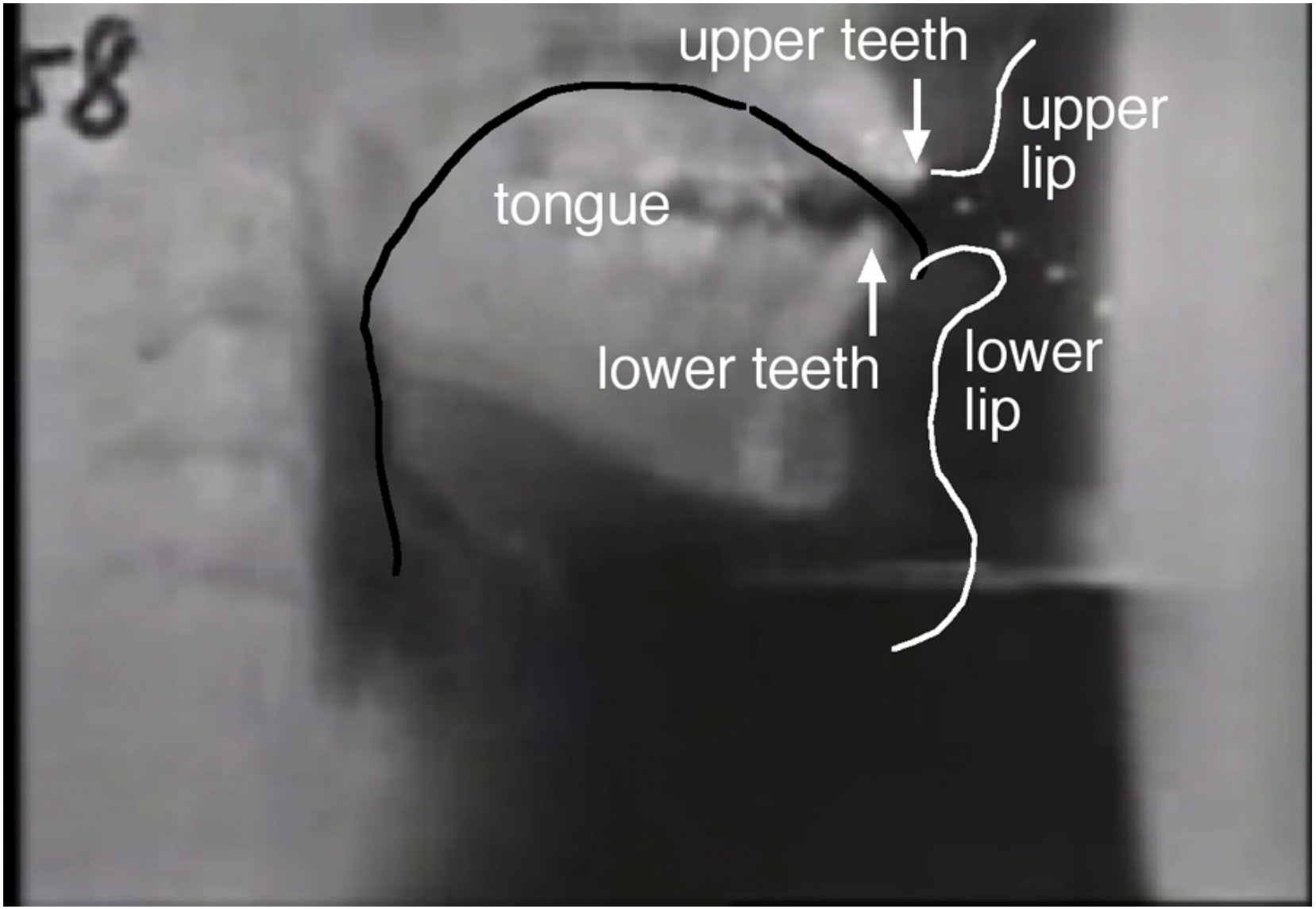
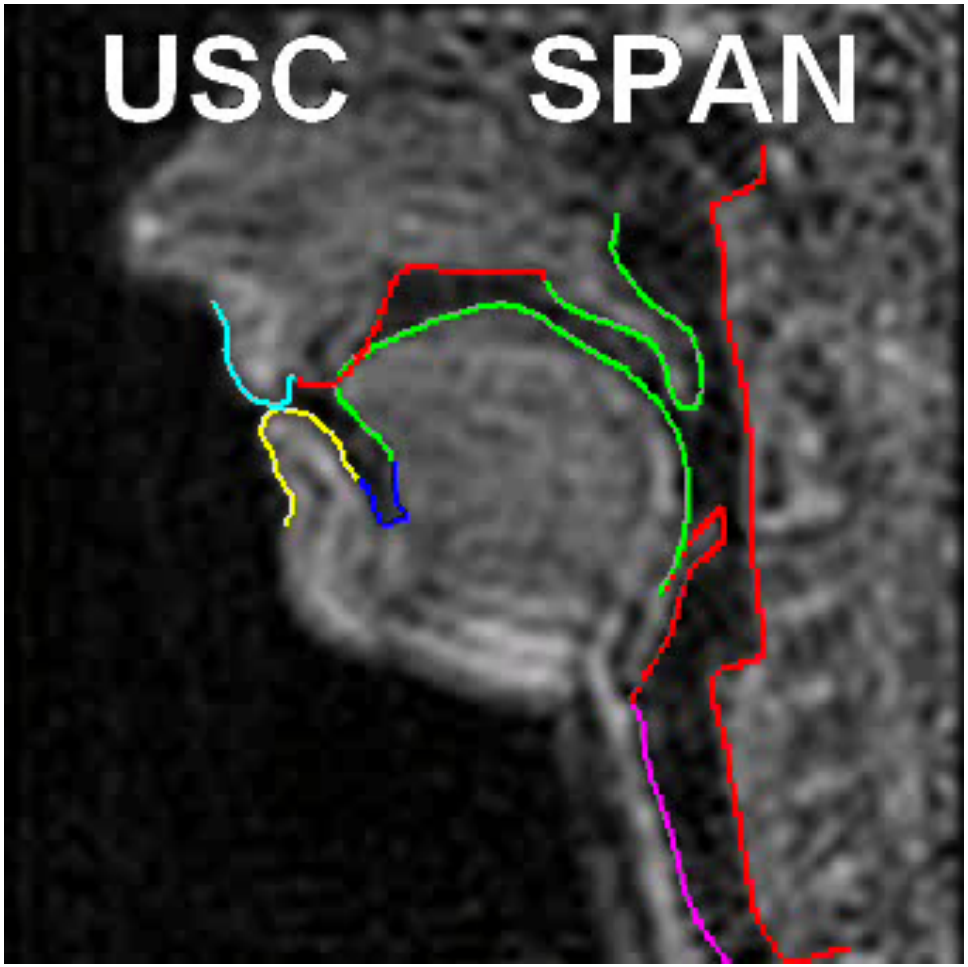


Figure of Ken Stevens, labels from Peter Ladefoged's web site

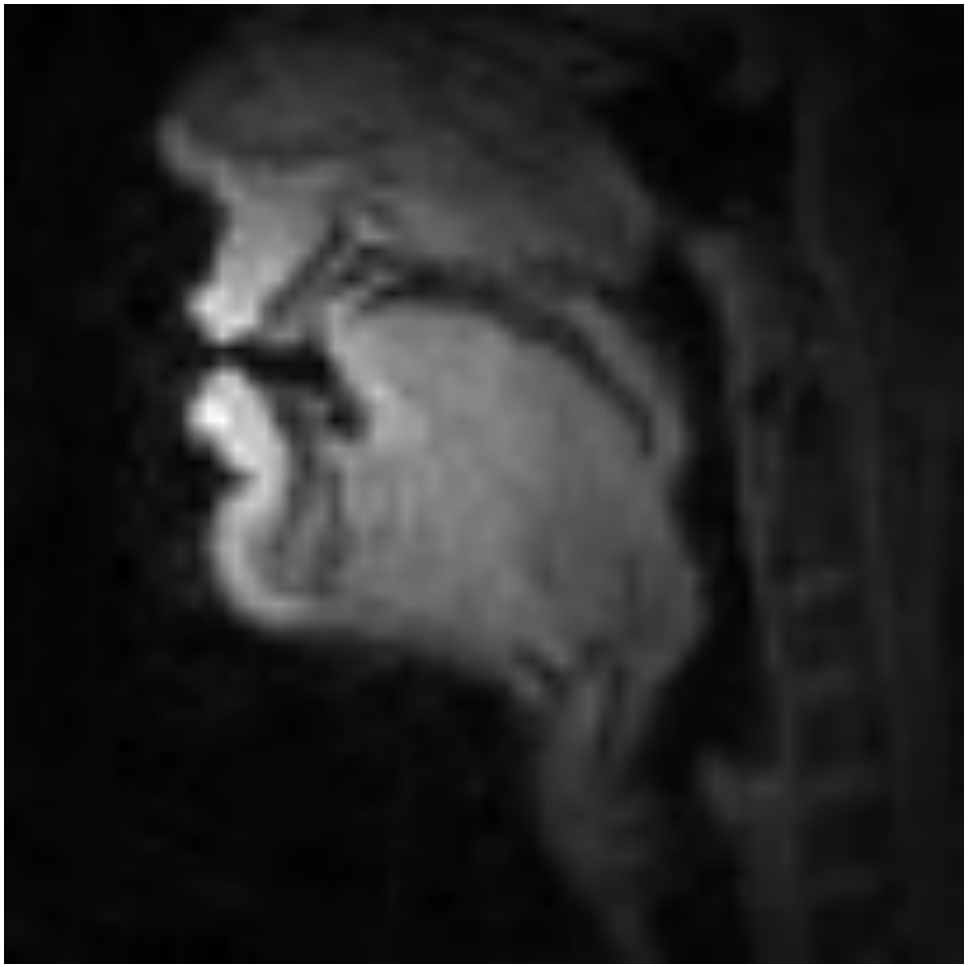
USC's SAIL Lab

Shri Narayanan




USC
SPAN

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)

Text from slides by Sharon Rose UCSD LING 111 handout

The larynx, external structure, from front

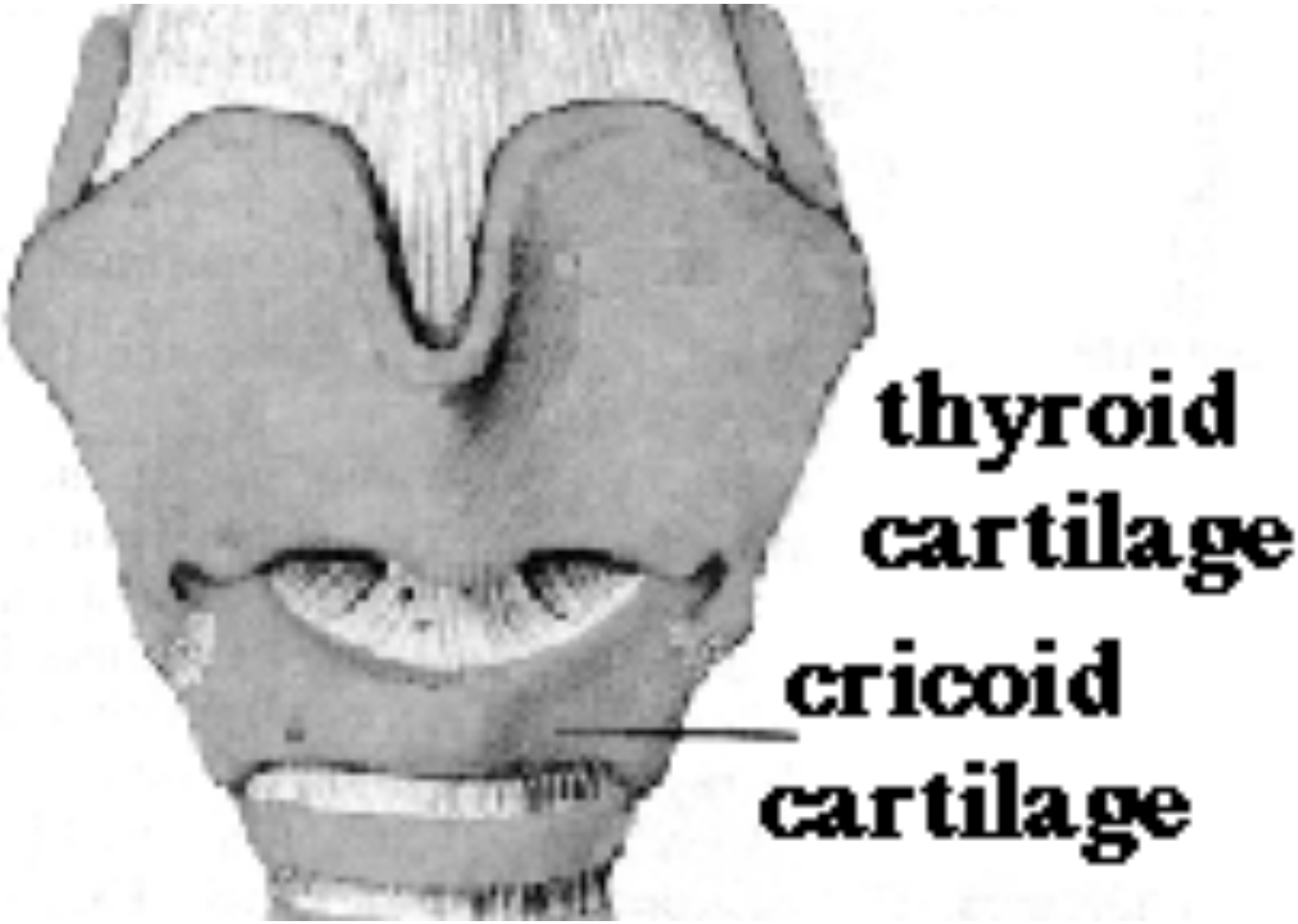


Figure thnx to John Coleman!!

Vertical slice through larynx, as seen from back

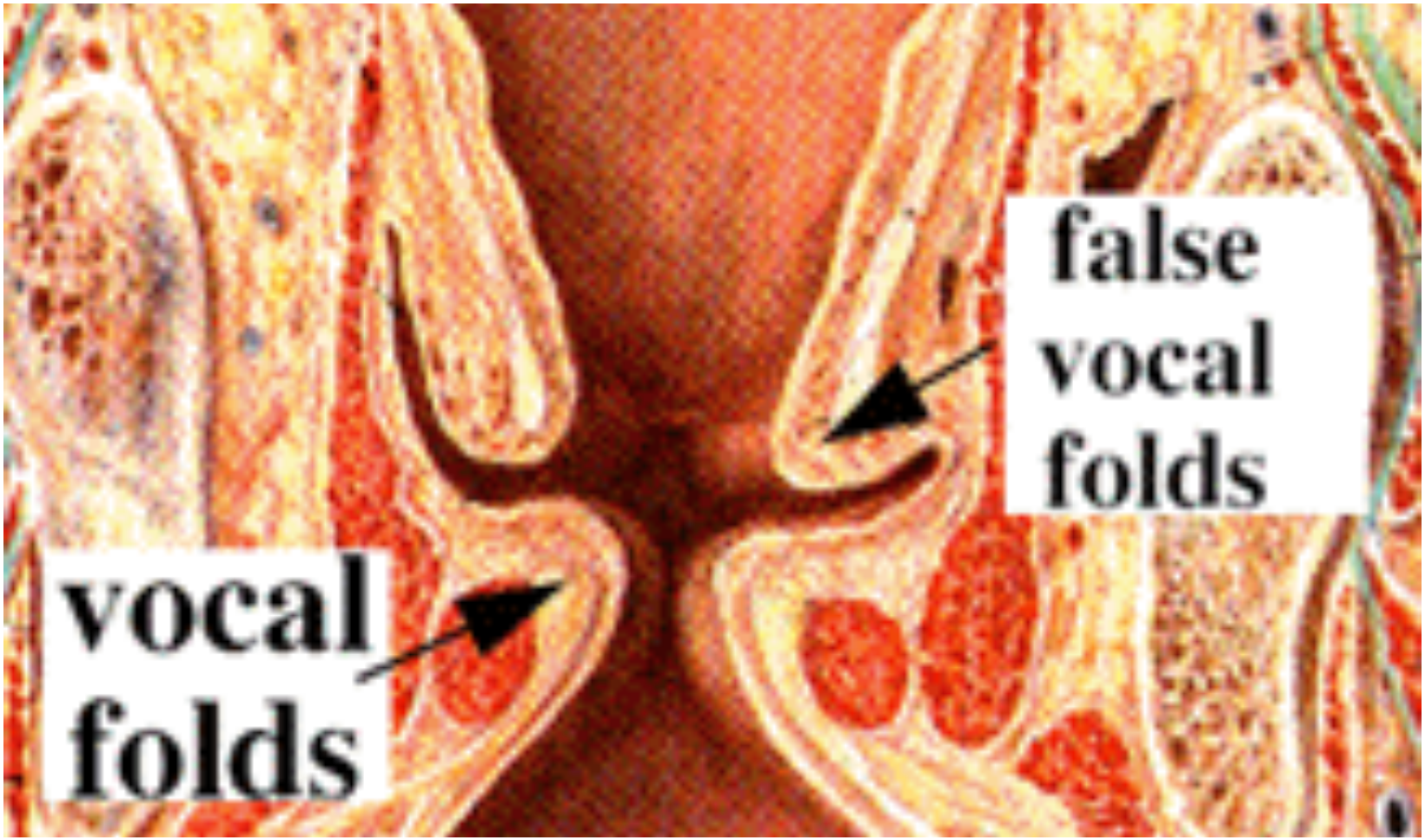
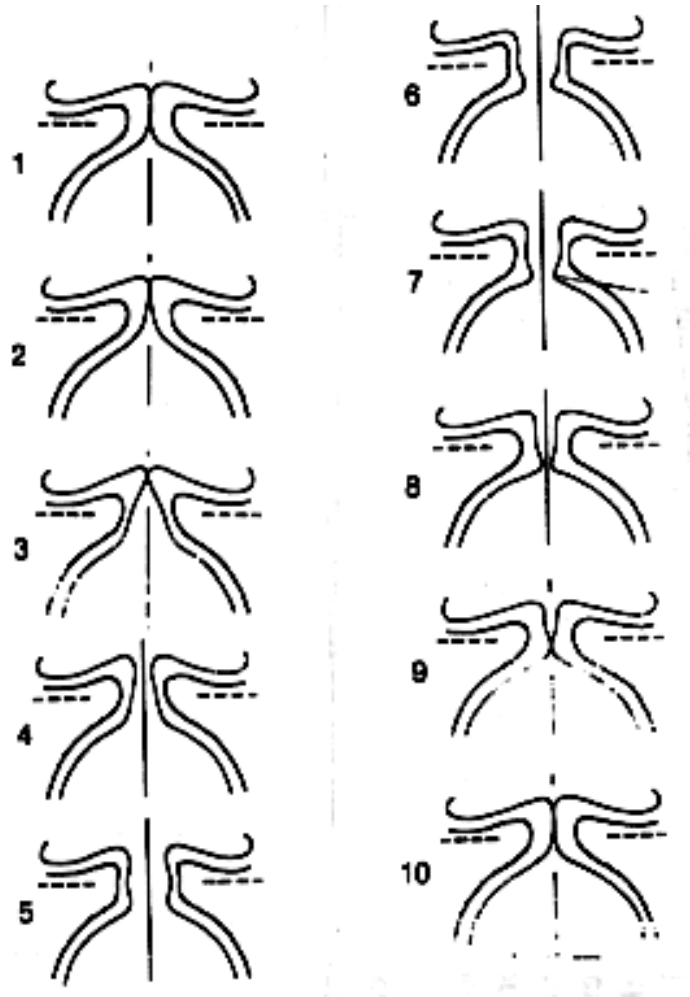


Figure thnx to John Coleman!!

Voicing:



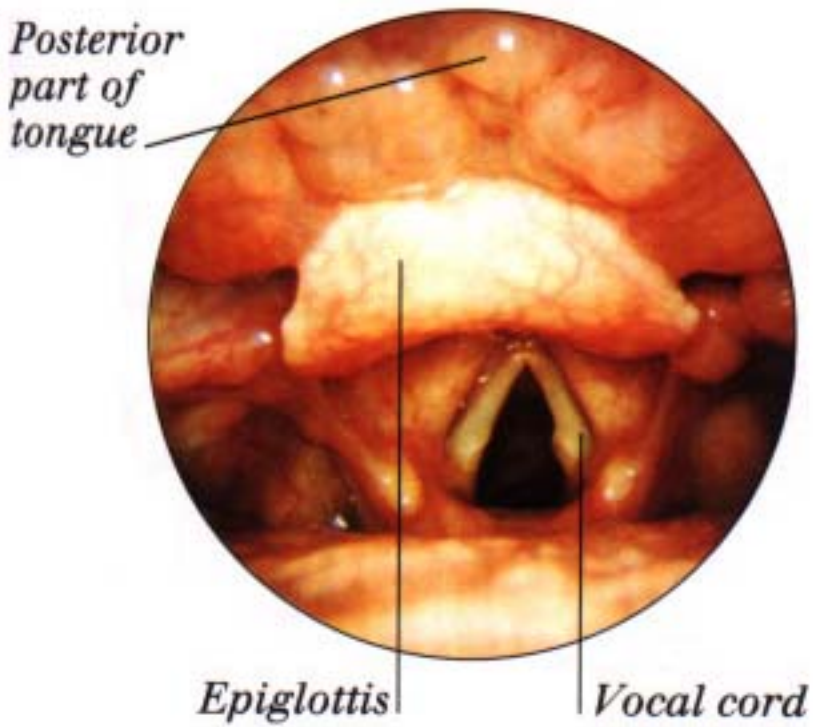
- Air comes up from lungs
- Forces its way through vocal cords, pushing open (2,3,4)
- This causes air pressure in glottis to fall, since:
  - when gas runs through constricted passage, its velocity increases (**Venturi tube effect**)
  - this increase in velocity results in a drop in pressure (**Bernoulli 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)

Text adapted from John Coleman

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

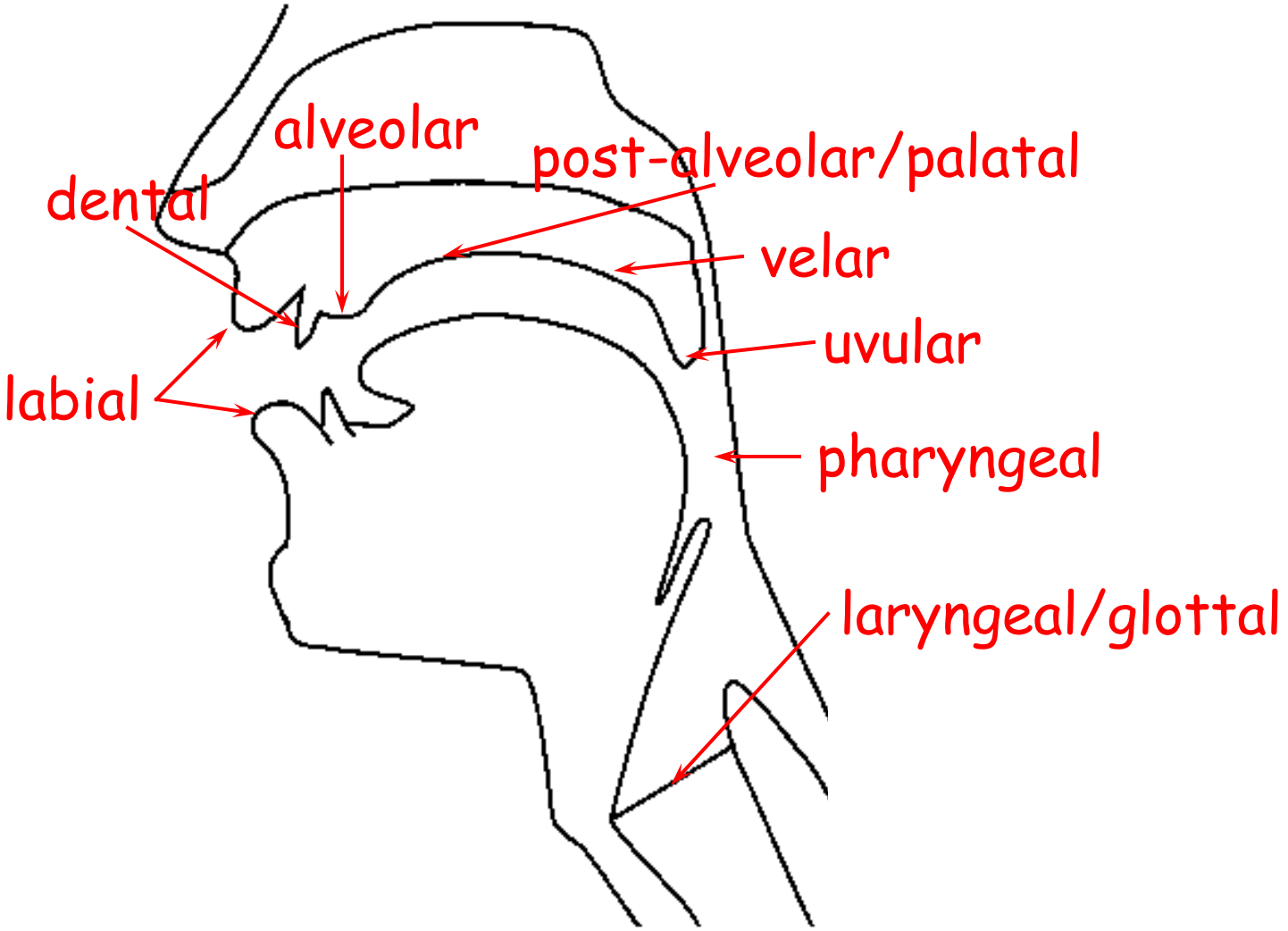
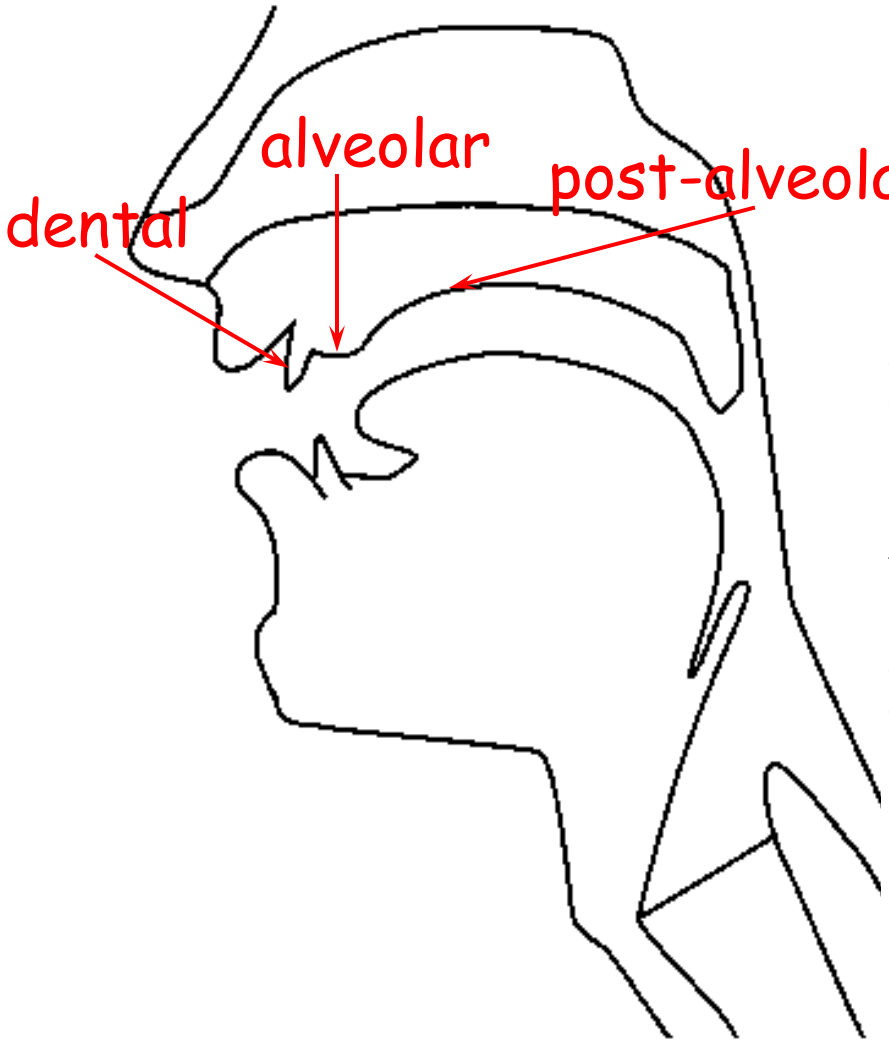


Figure thanks to Jennifer Venditti

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

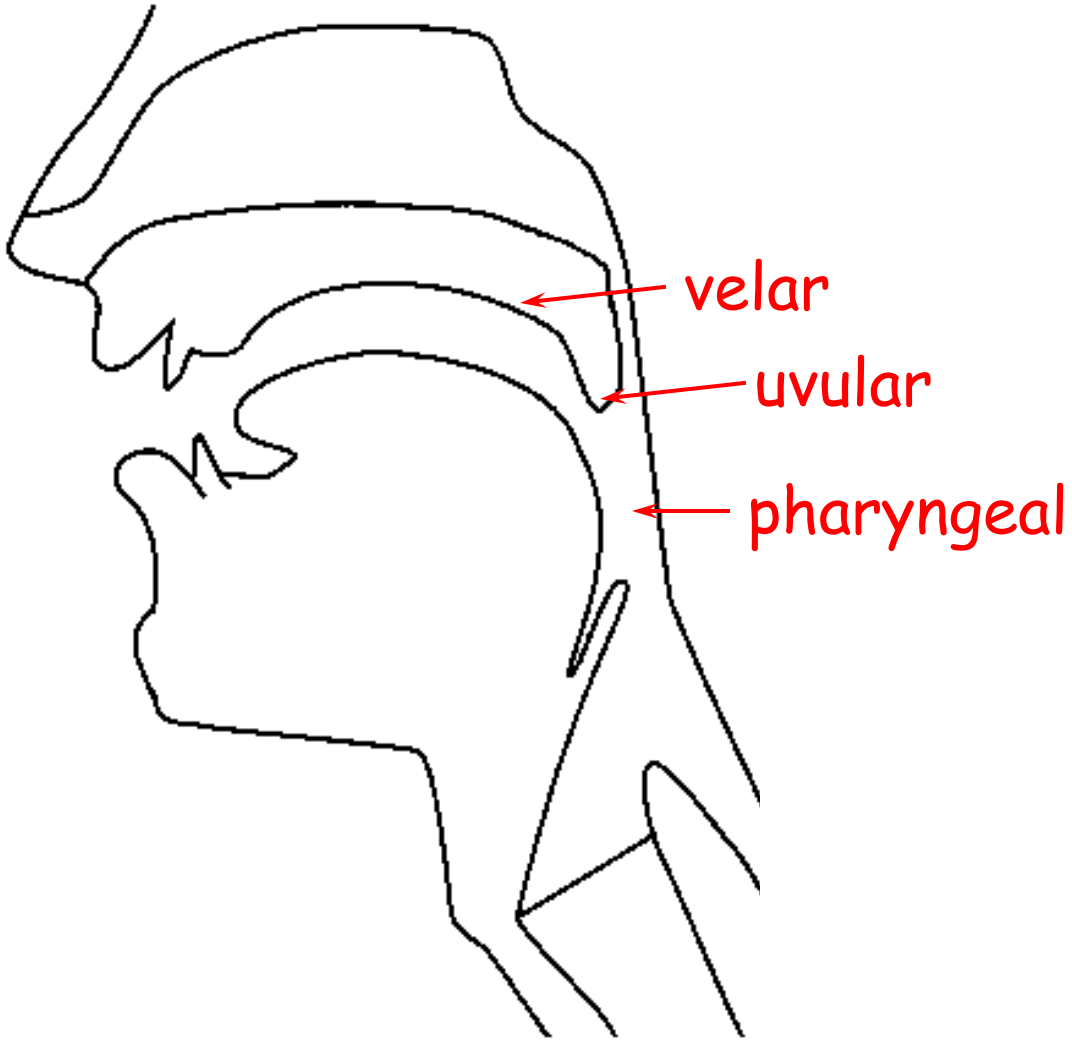
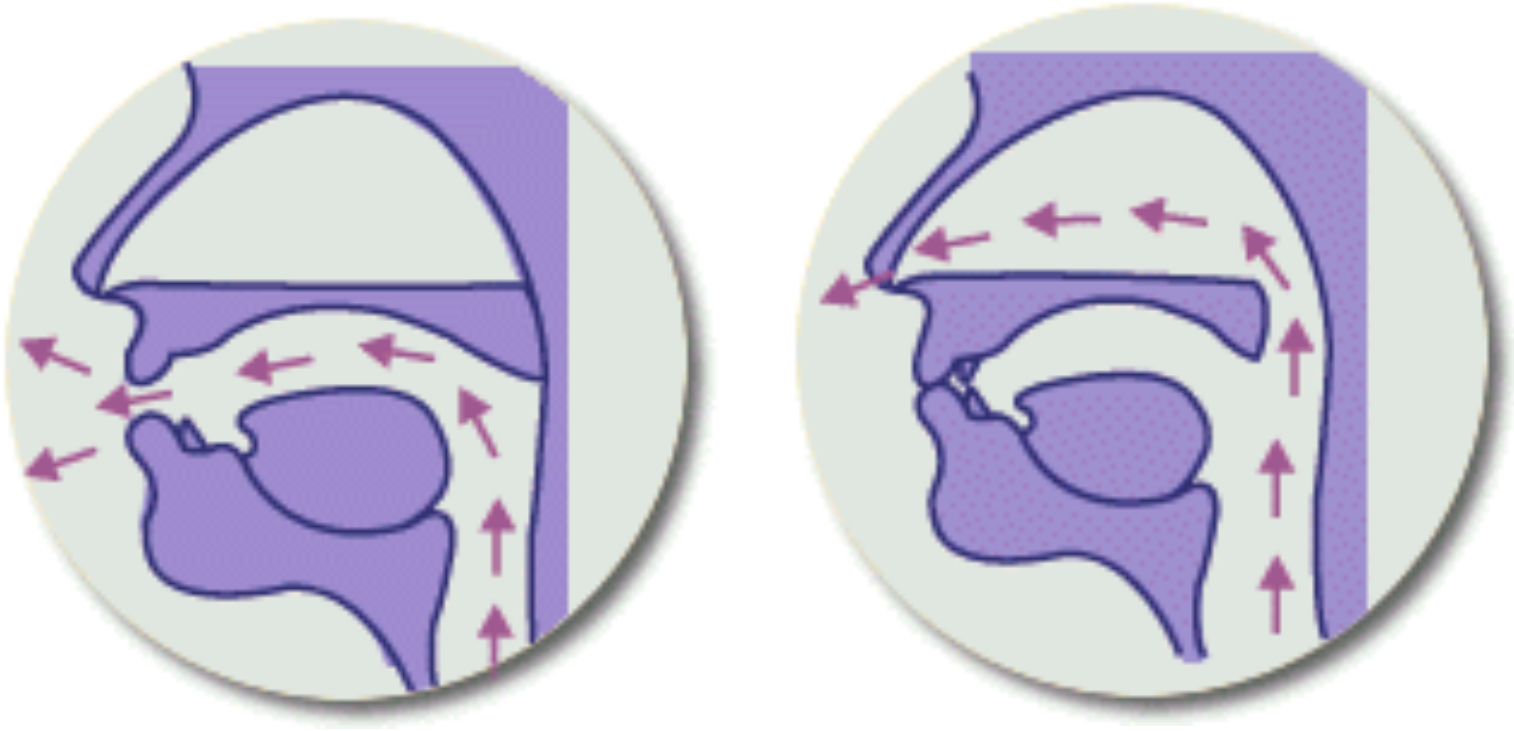


Figure thanks to Jennifer Venditti

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

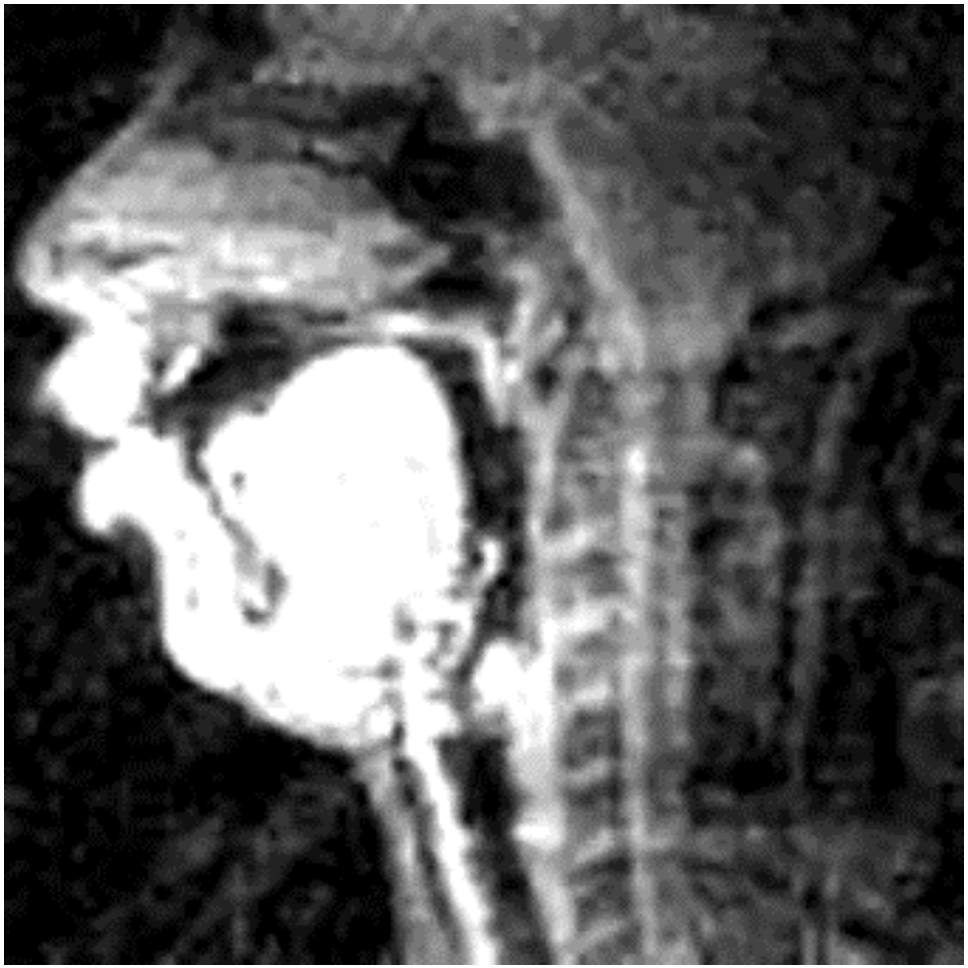
Articulatory parameters for English consonants (in ARPAbet)

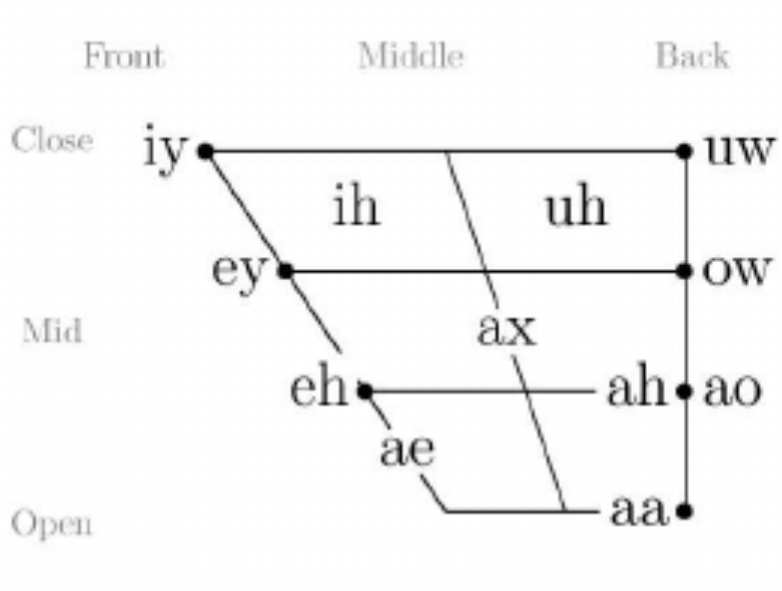
MANNER OF ARTICULATION	PLACE OF ARTICULATION													
	bilabial		labio-dental		inter-dental		alveolar		palatal		velar		glottal	
stop	p	b					t	d			k	g	q	
fric.			f	v	th	dh	s	z	sh	zh			h	
affric.									ch	jh				
nasal		m						n				ng		
approx		w						l/r		y				
flap							dx							

VOICING: voiceless | voiced

Table from Jennifer Venditti

Tongue position for vowels





Vowels

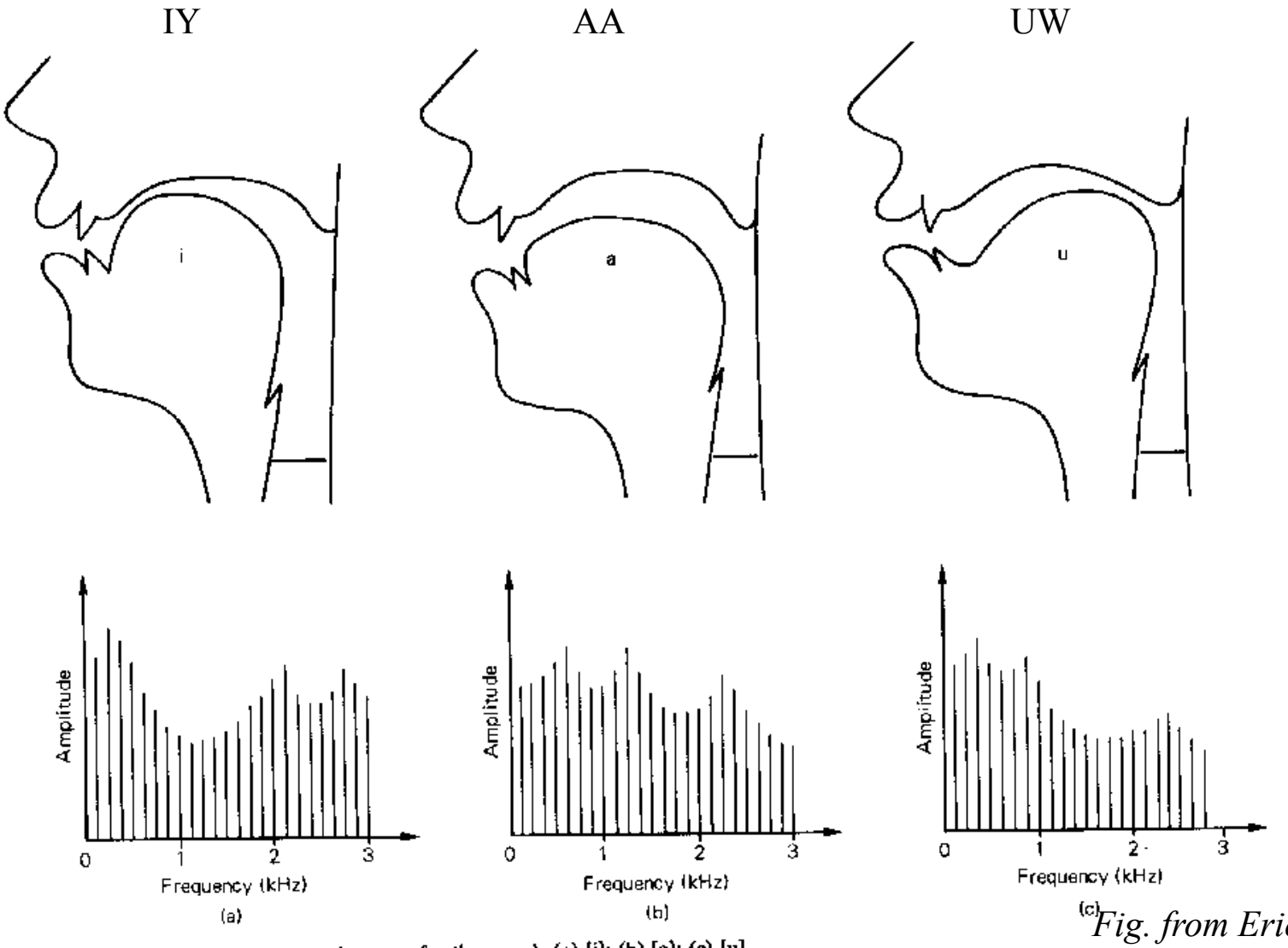
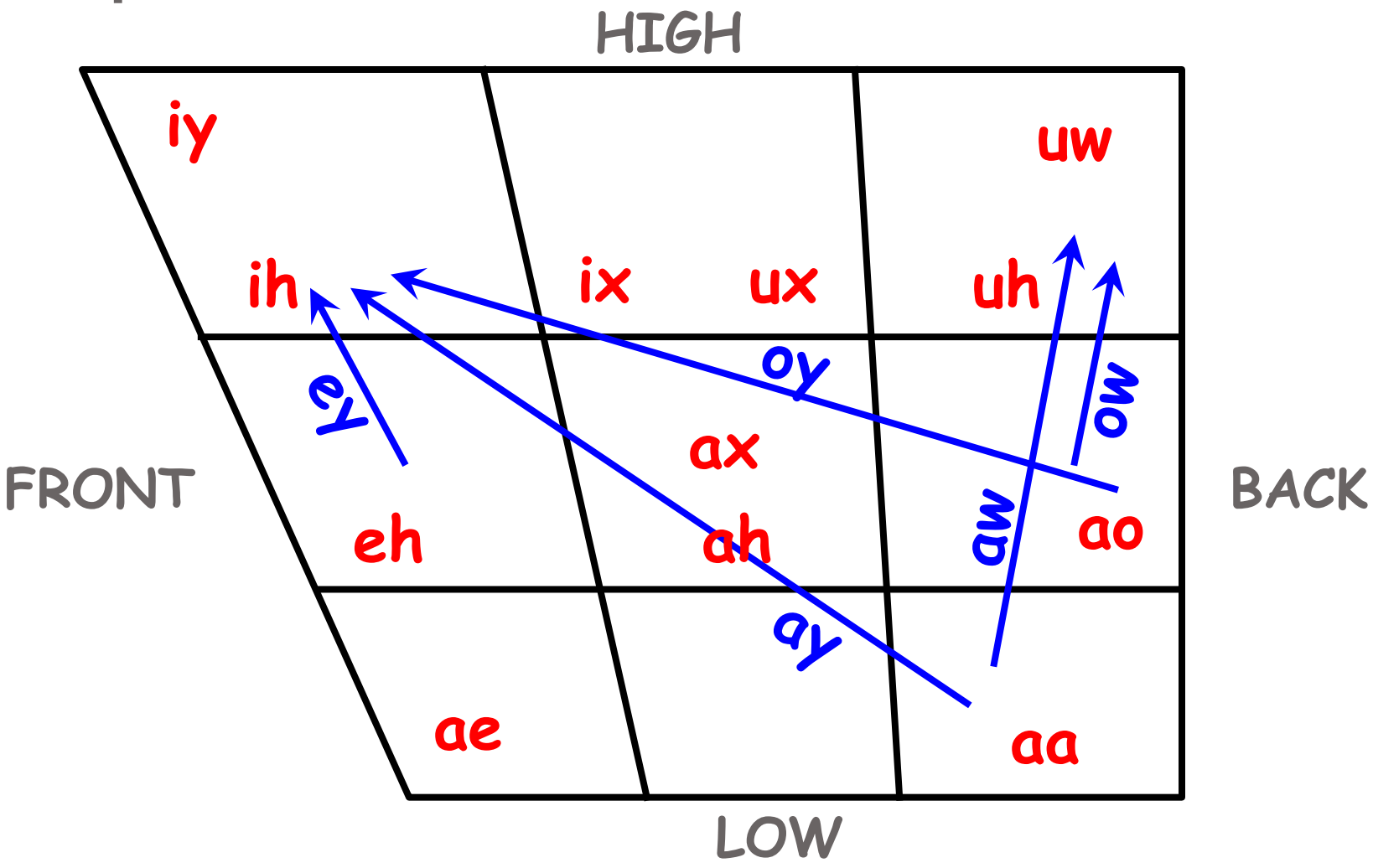


Fig. from Eric Keller

American English Vowel Space



Red: Vowels, Blue: Dipthongs

Figure from Jennifer Venditti

[iy] vs. [uw]

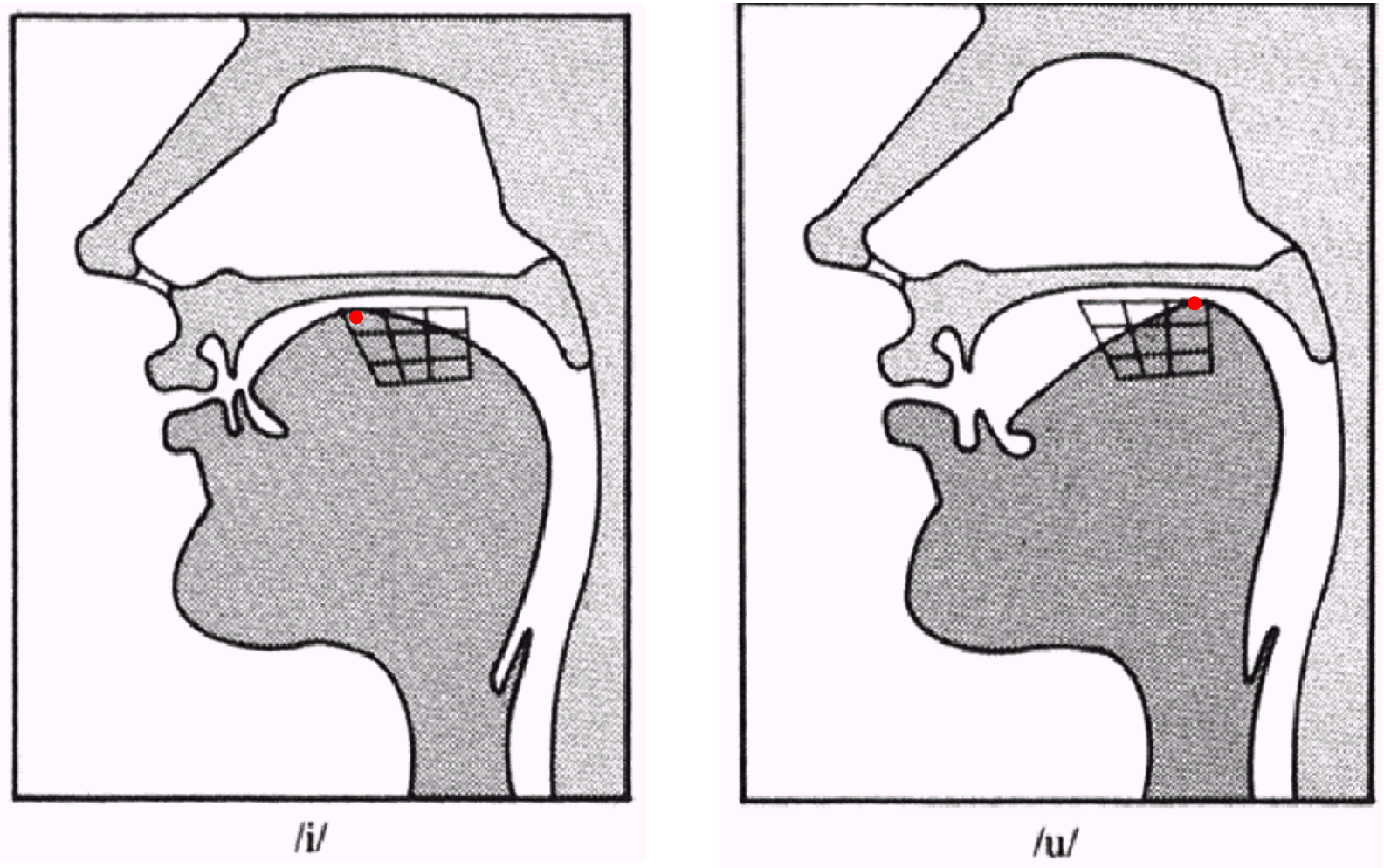


Figure from Jennifer Venditti, from a lecture given by Rochelle Newman

[ae] vs. [aa]

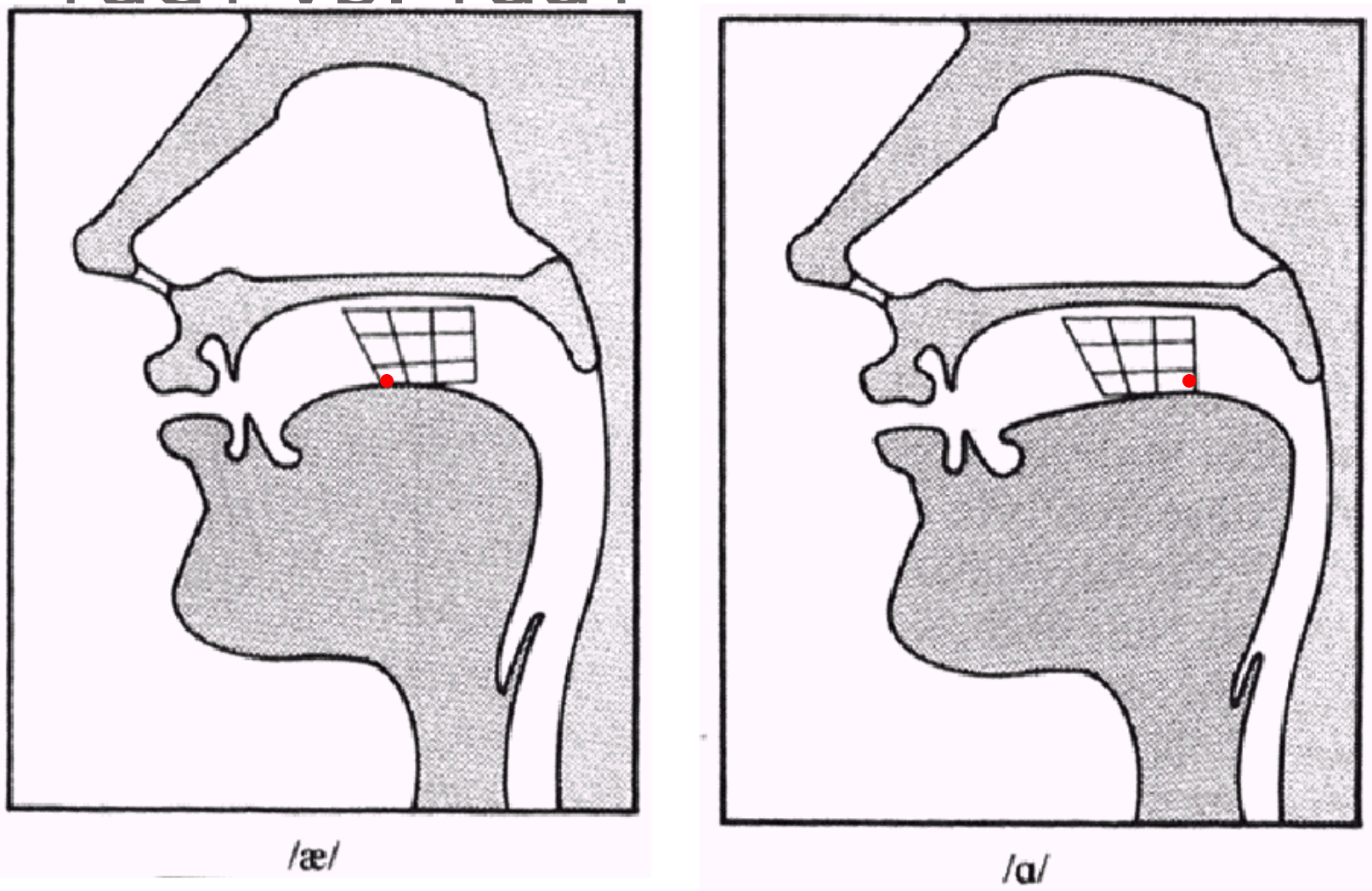
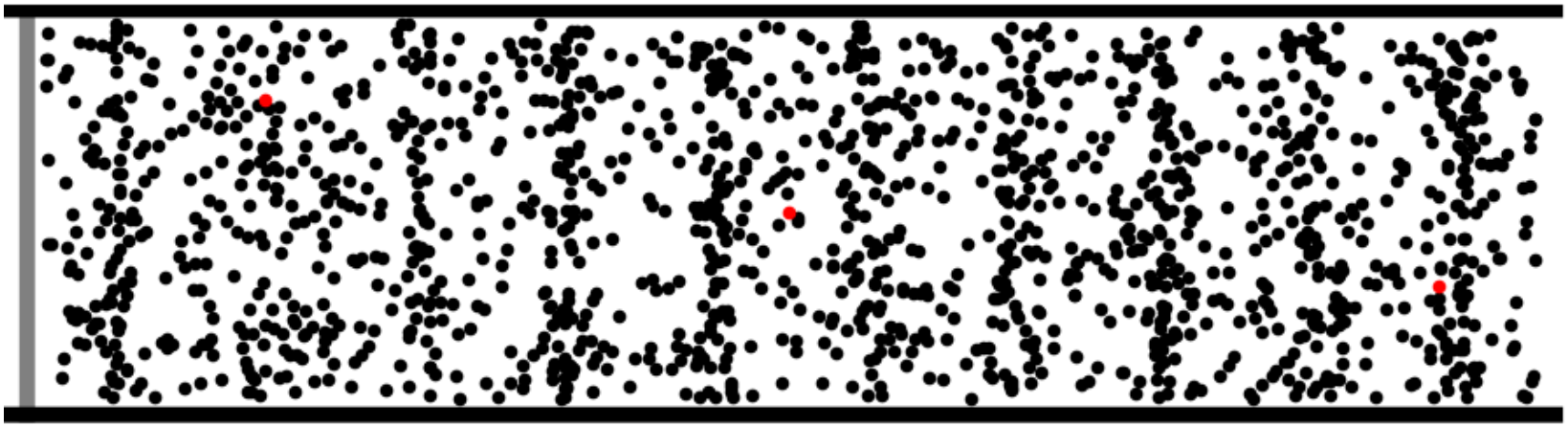


Figure from Jennifer Venditti, from a lecture given by Rochelle Newman

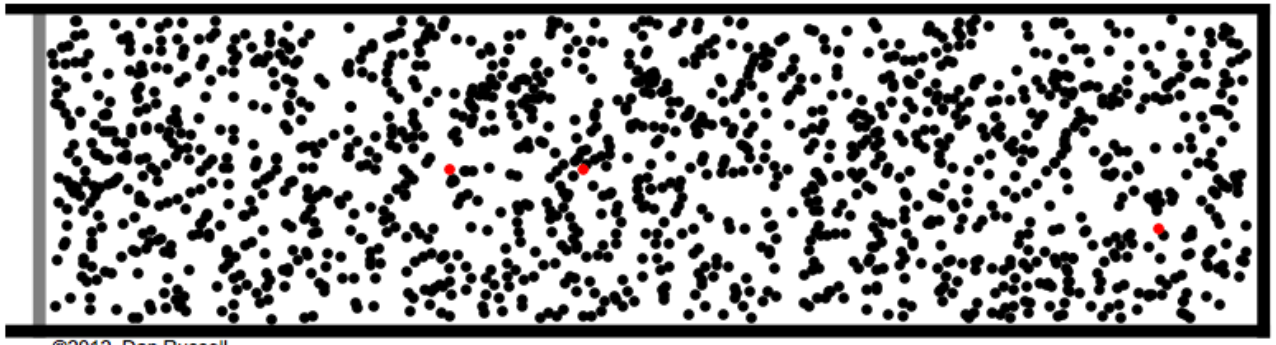
Where to go for more info

- 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
- Jennifer Smith's resource page
  - http://www.unc.edu/~jlsmith/pht-url.html

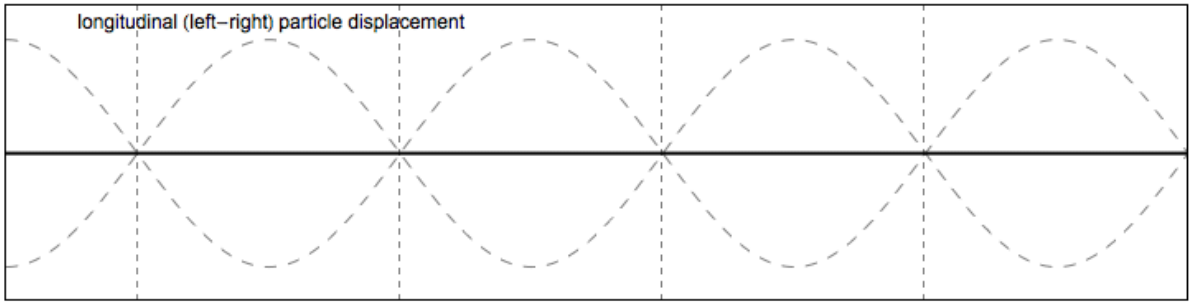
Sound waves are longitudinal waves



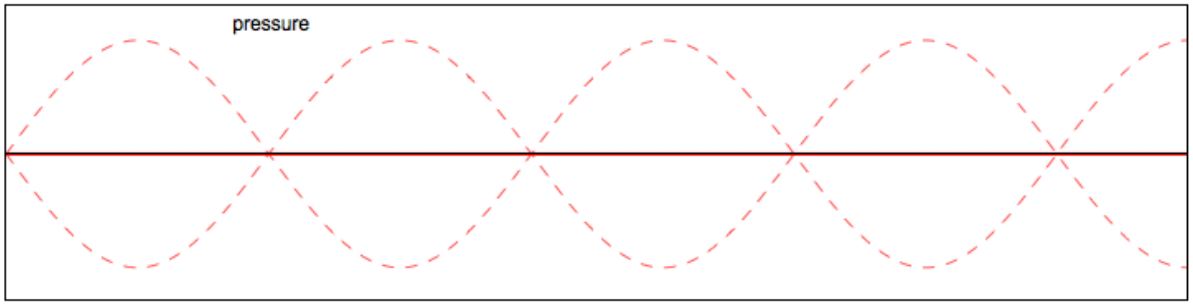
Dan Rusell Figure



particle dispacment

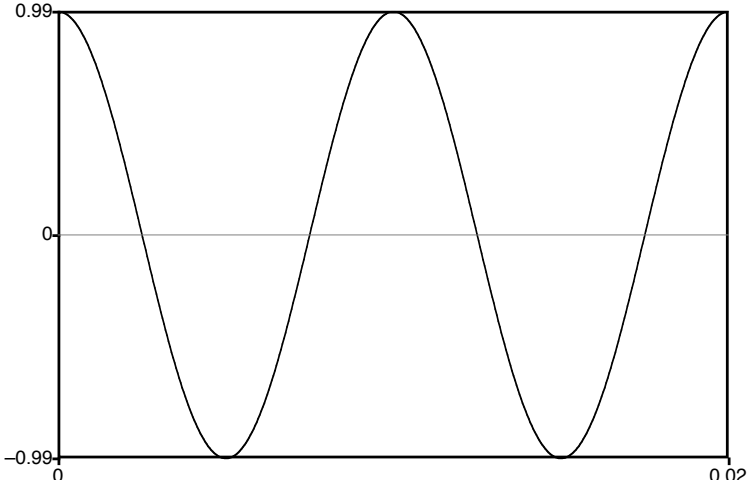


pressure



Dan Rusell Figure

Remember High School Physics Simple Period Waves (sine waves)



- Characterized by:
  - period: T
  - amplitude A
  - phase ϕ
- Fundamental frequency

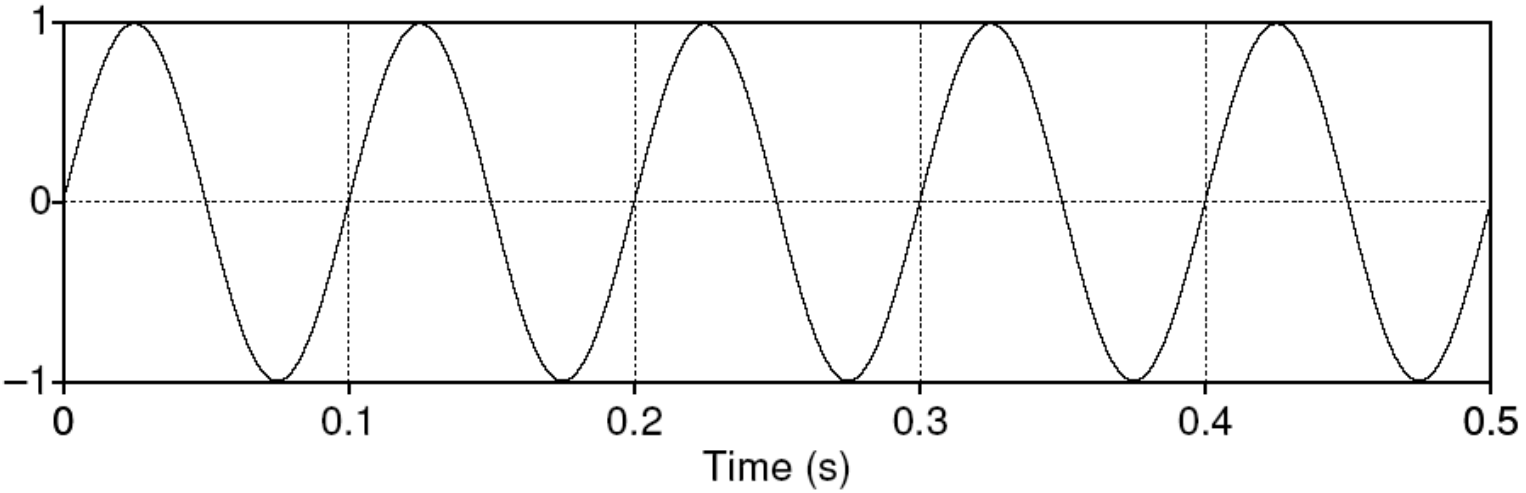
in cycles per second, or Hz

  - $F_0=1/T$

1 cycle

To listen to sine waves:
http://www.szynalski.com/tone-generator/

Simple periodic waves

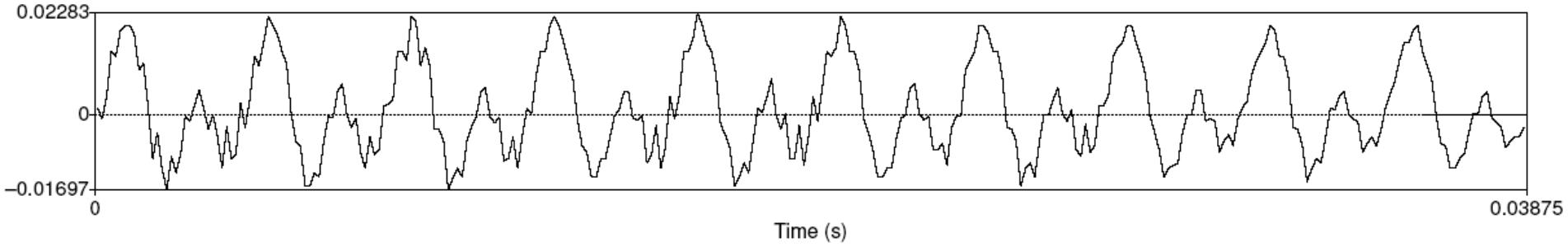


The frequency of a wave:

5 cycles in .5 seconds = 10 cycles/second = 10 Hz

Amplitude: 1

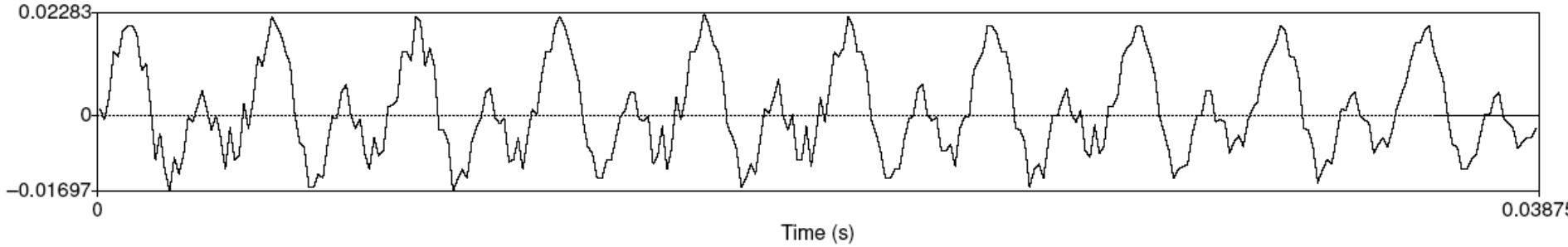
Speech sound waves



- X axis: time.
- Y axis:
  - Amplitude = air pressure at that time
    - +: compression
    - 0: normal air pressure,
    - -: rarefaction

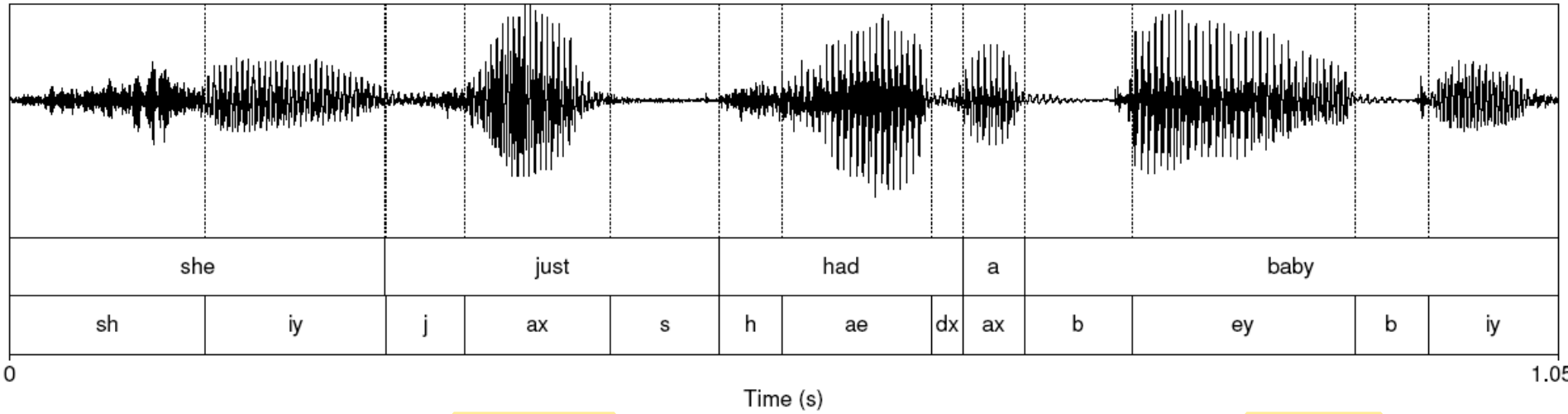
Back to waves: Fundamental frequency

- Waveform of the vowel [iy]



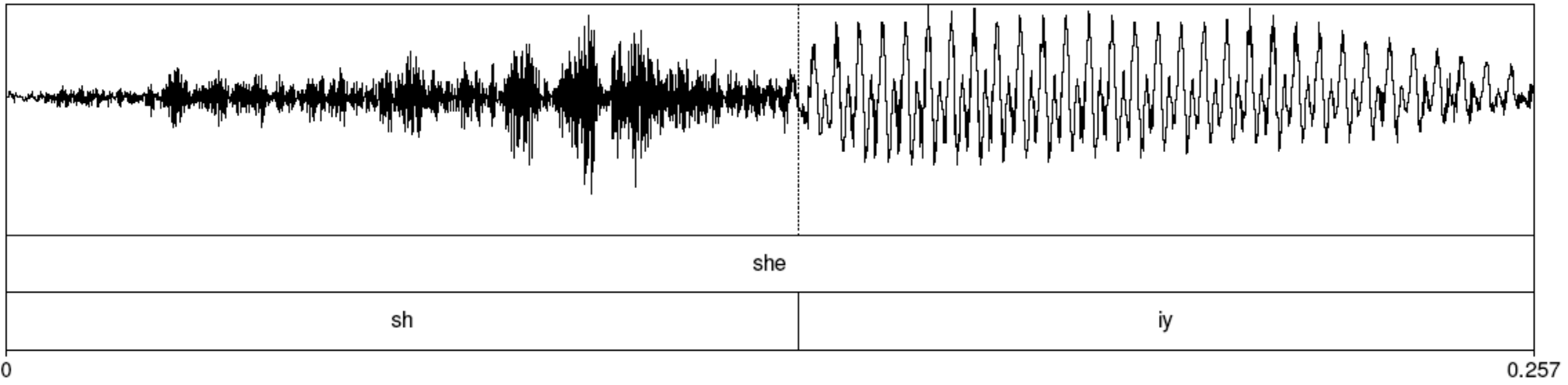
- Frequency: 10 repetitions / .03875 seconds = 258 Hz
- This is speed that vocal folds move, hence voicing
- Each peak corresponds to an opening of the vocal folds
- The low frequency of the complex wave is called the fundamental frequency of the wave or F0

She just had a baby

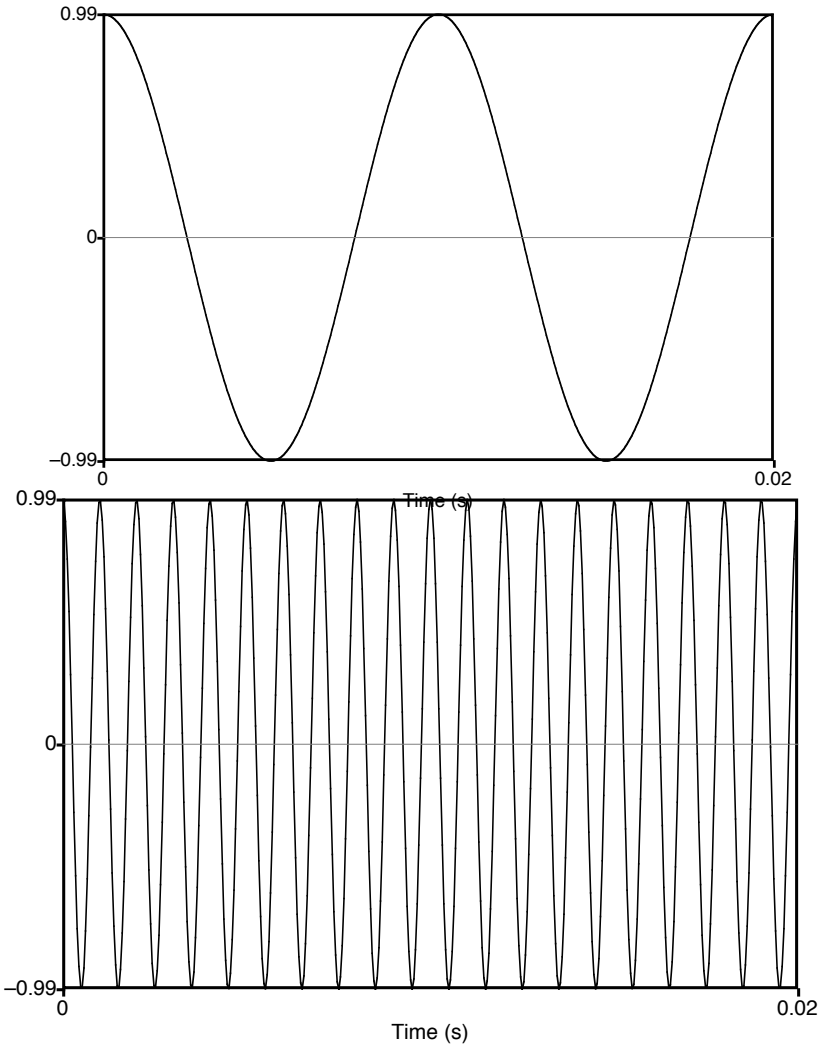


- Note that vowels all have regular amplitude peaks
- Stop consonant
  - Closure followed by release
  - Notice the silence followed by slight bursts of emphasis: very clear for [b] of “baby”
- Fricative: noisy. [sh] of “she” at beginning

Fricative



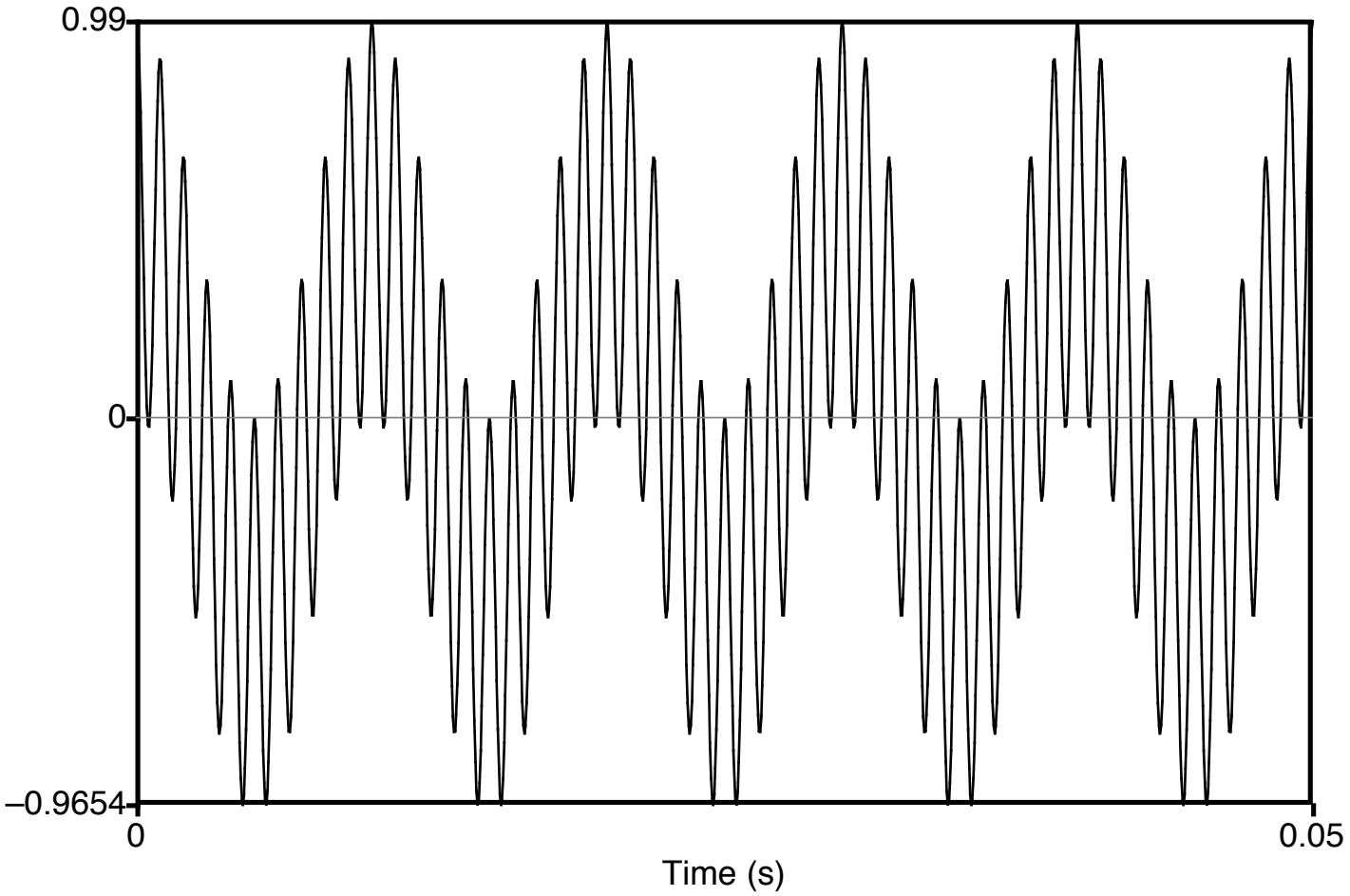
Back to freshman physics: Waves have different frequencies



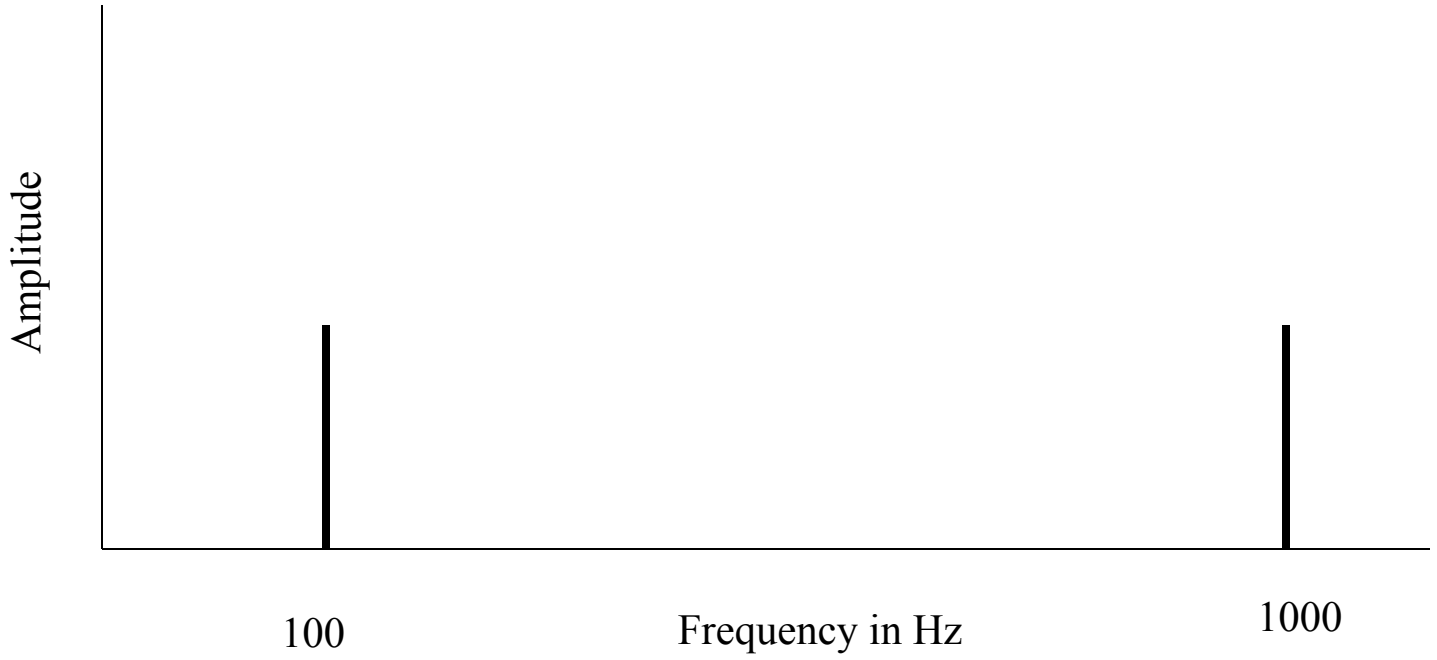
100 Hz

1000 Hz

Complex waves: Adding a 100 Hz and 1000 Hz wave together



Spectrum

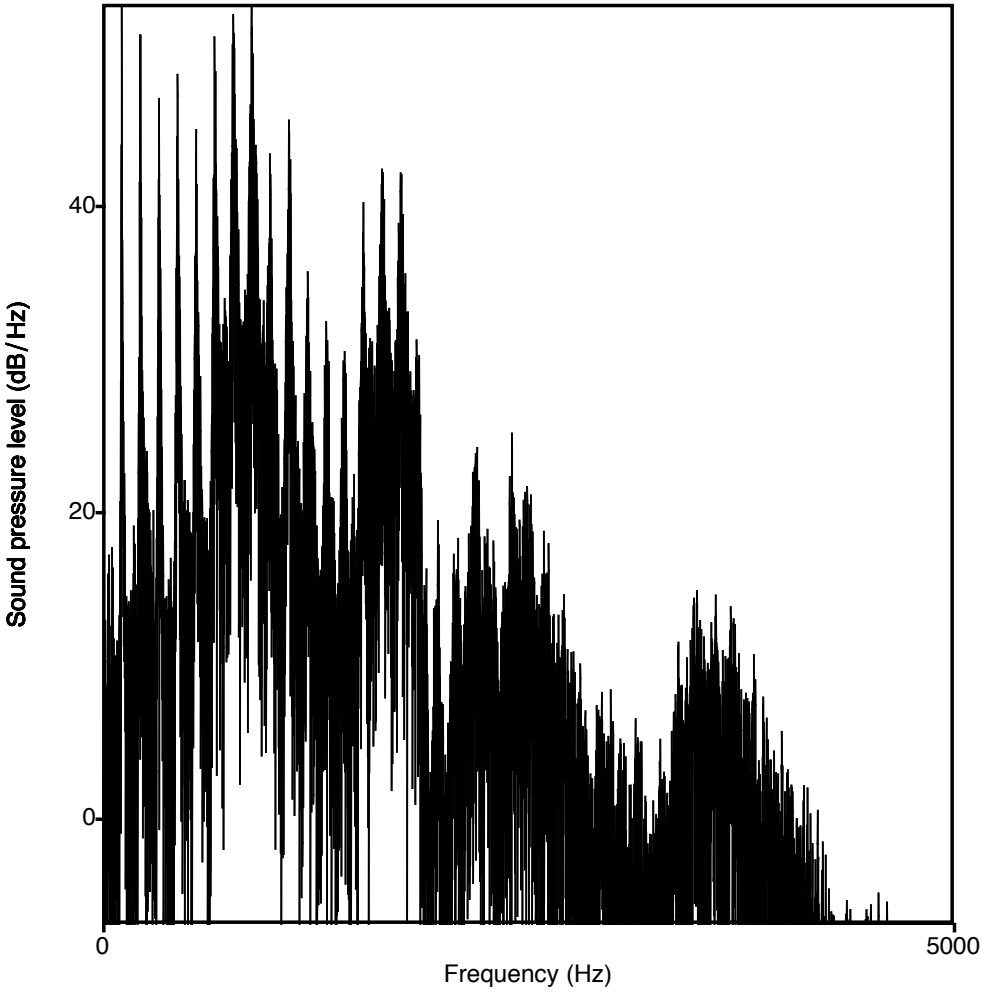


Frequency components (100 and 1000 Hz) on x-axis

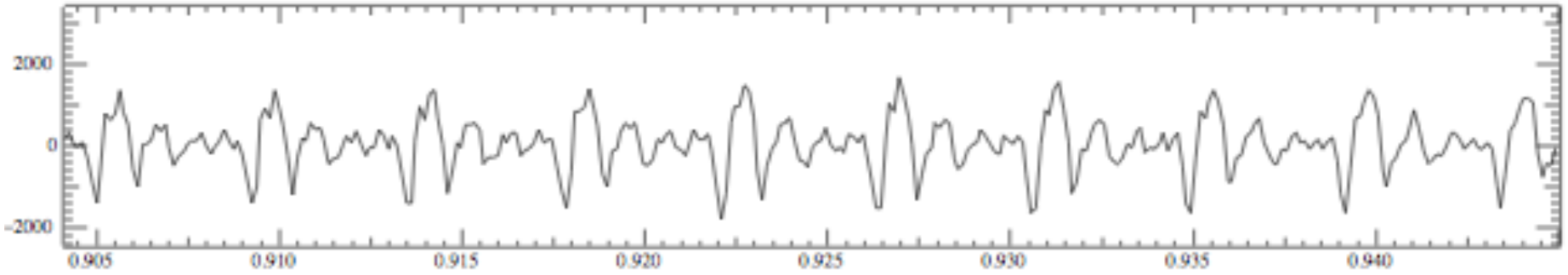
Spectra continued

- Fourier analysis: any wave can be represented as the (infinite) sum of sine waves of different frequencies (amplitude, phase)

Spectrum of one instant in an actual soundwave: many components across frequency range



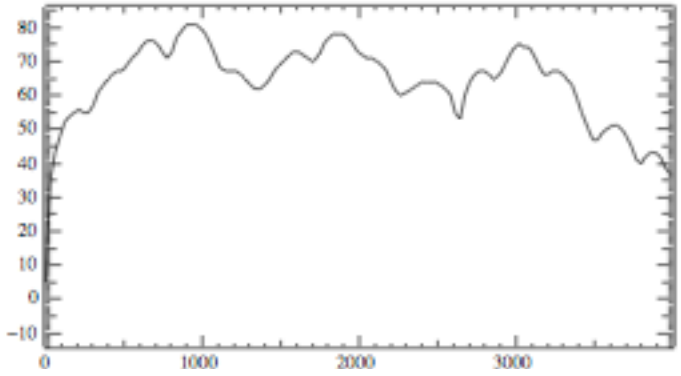
Part of [ae] waveform from “had”



- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves

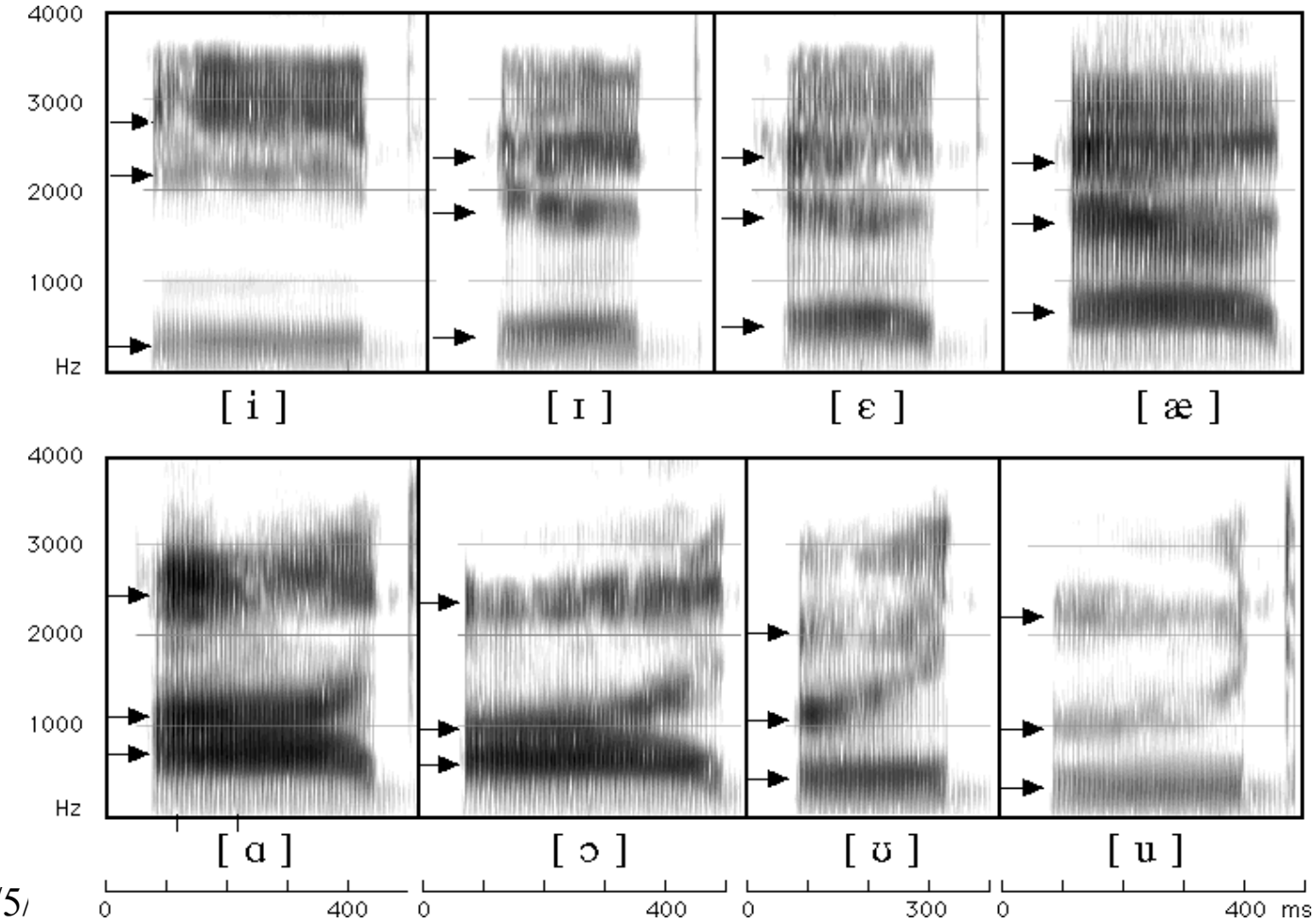
Back to spectrum

- Spectrum represents these freq components
- Computed by Fourier transform



- x-axis shows frequency, y-axis shows magnitude (in decibels)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

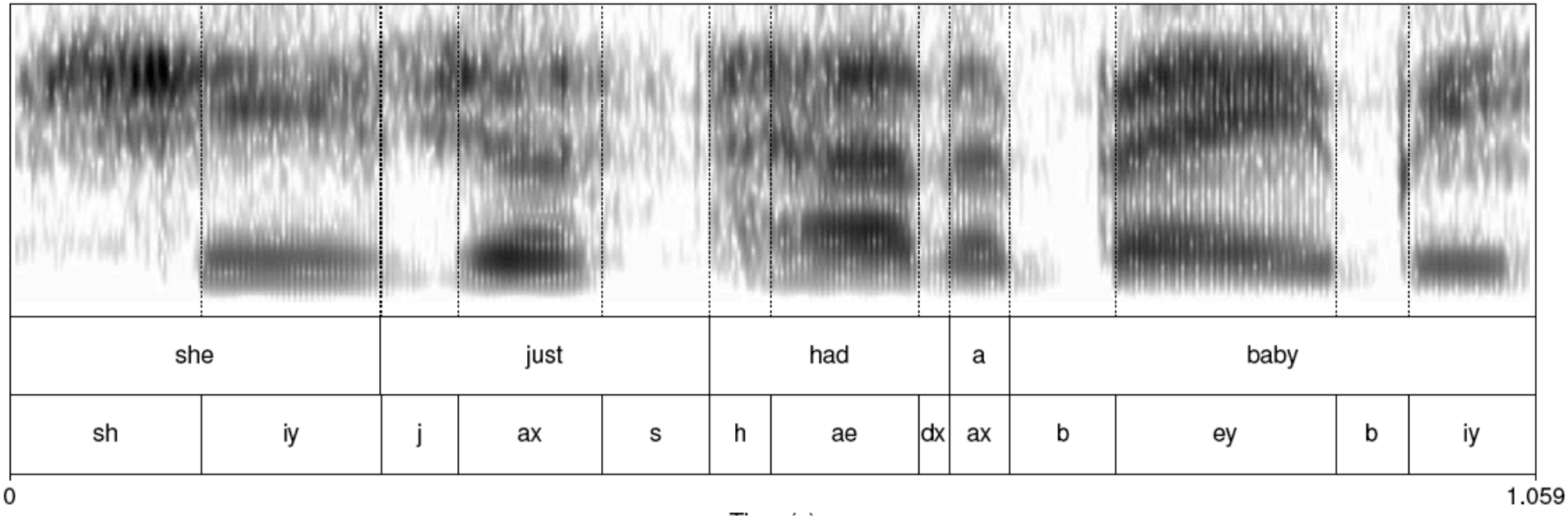
Seeing formants: the spectrogram



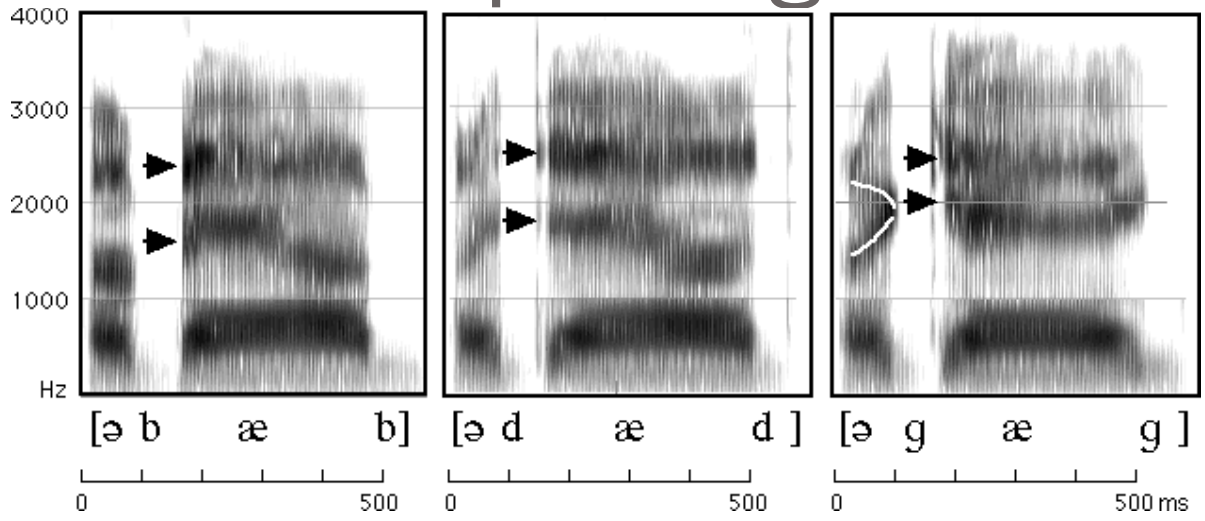
Formants

- Vowels largely distinguished by 2 characteristic pitches.
- One of them (the higher of the two) goes downward throughout the series iy ih eh ae aa ao ou u
- The other goes up for the first four vowels and then down for the next four.
- These are called "formants" of the vowels, lower is 1st formant, higher is 2nd formant.

Spectrogram: spectrum + time dimension



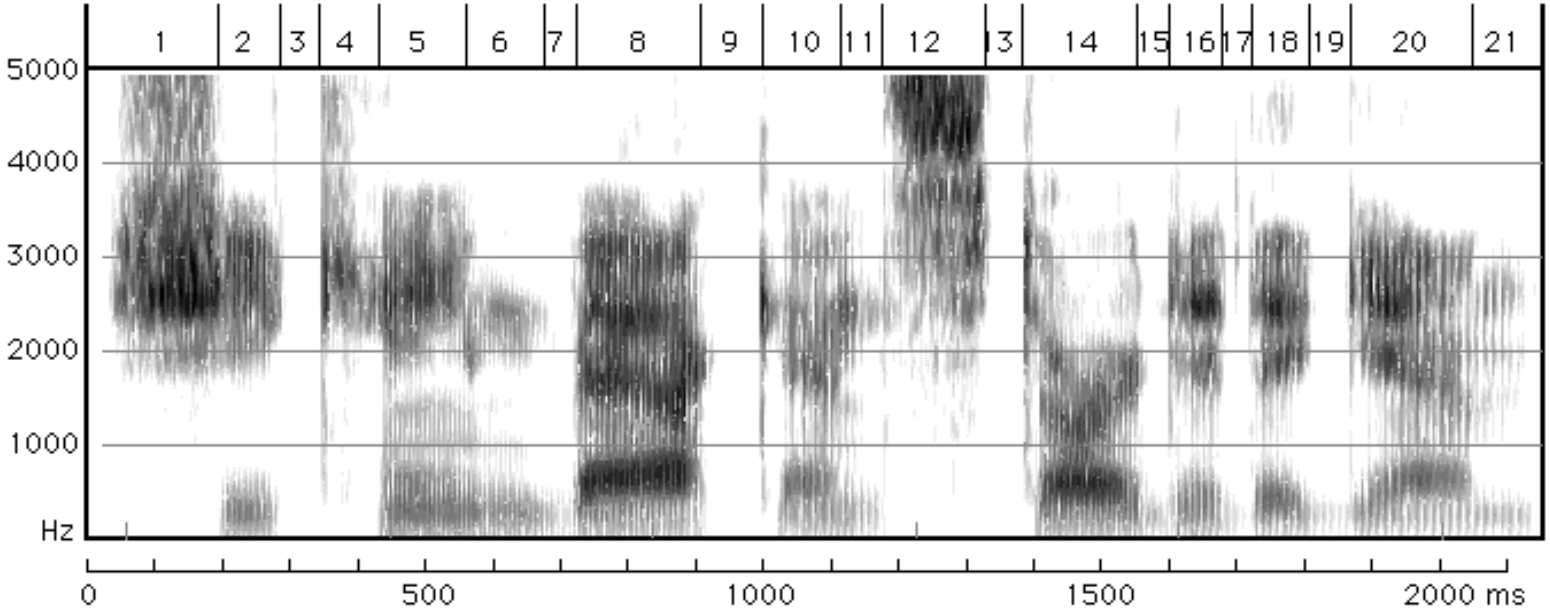
How to read spectrograms



- 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

From Ladefoged "A Course in Phonetics"

She came back and started again



- 1. lots of high-freq energy
- 3. closure for k
- 4. burst of aspiration for k
- 5. ey vowel;faint 1100 Hz formant is nasalization
- 6. bilabial nasal
- 7. short b closure, voicing barely visible.
- 8. ae; note upward transitions after bilabial stop at beginning
- 9. note F2 and F3 coming together for "k"

From Ladefoged "A Course in Phonetics"

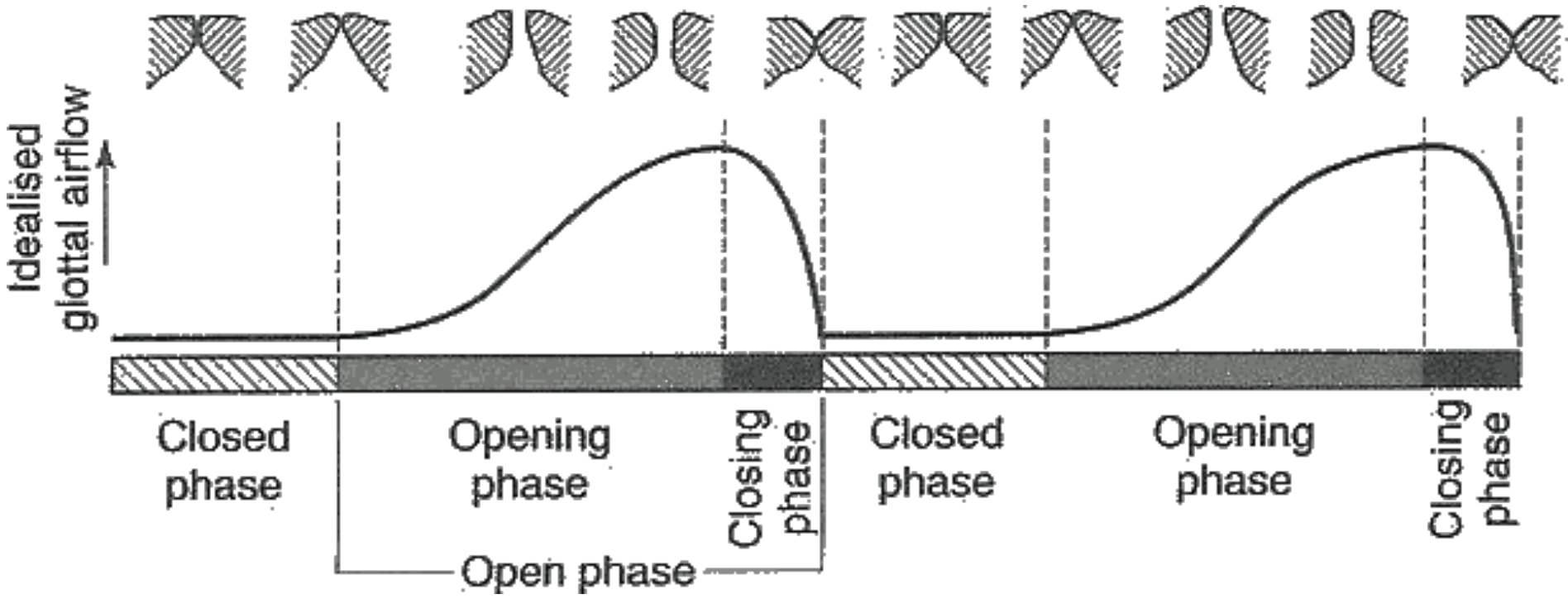
Praat example

- http://www.fon.hum.uva.nl/praat/

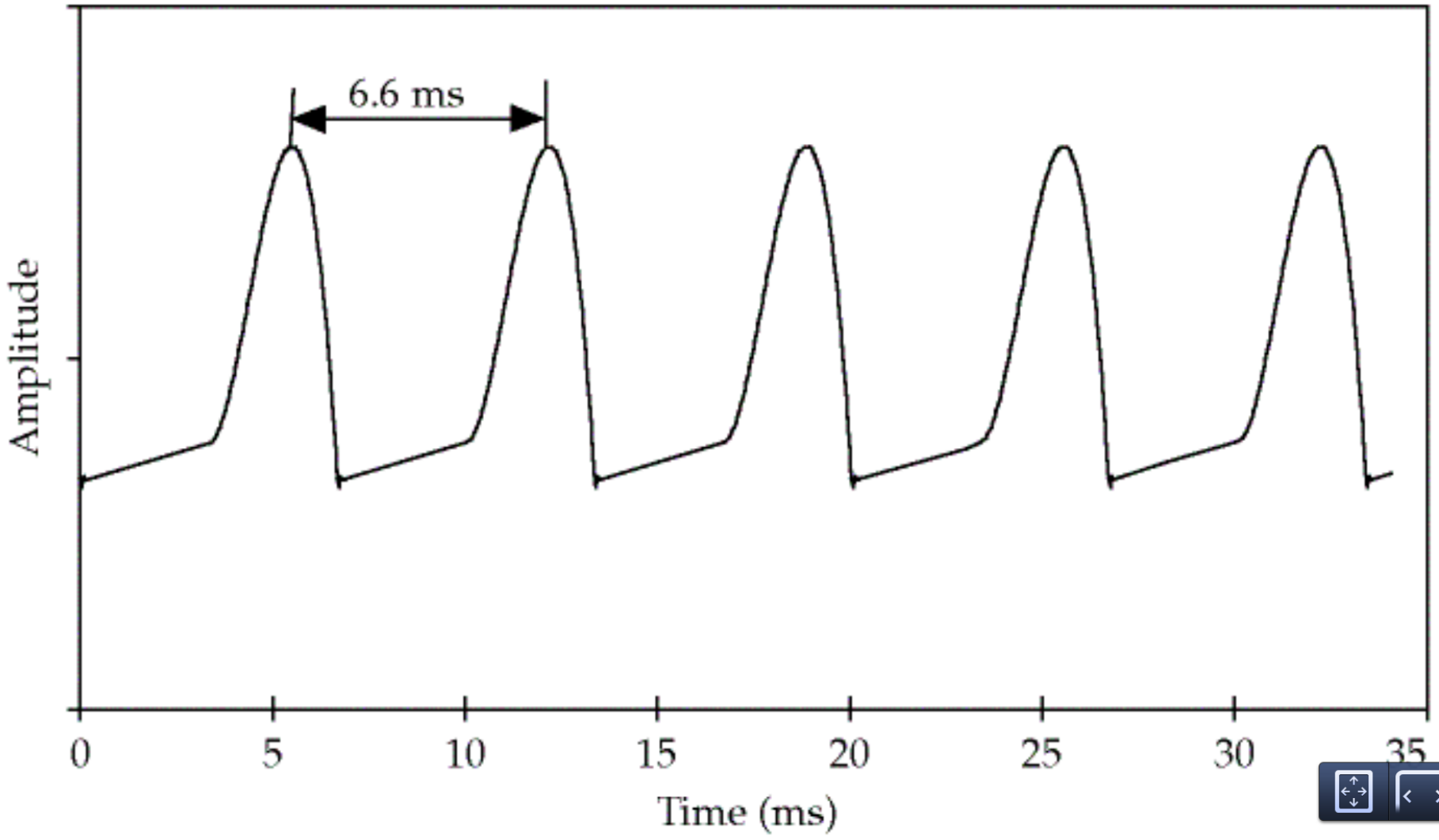
Different vowels have different formants

- Every time the vocal cords open and close, pulse of air from the lungs is sharp tap on air in vocal tract.
- Setting air in vocal cavity vibrating, producing different harmonics

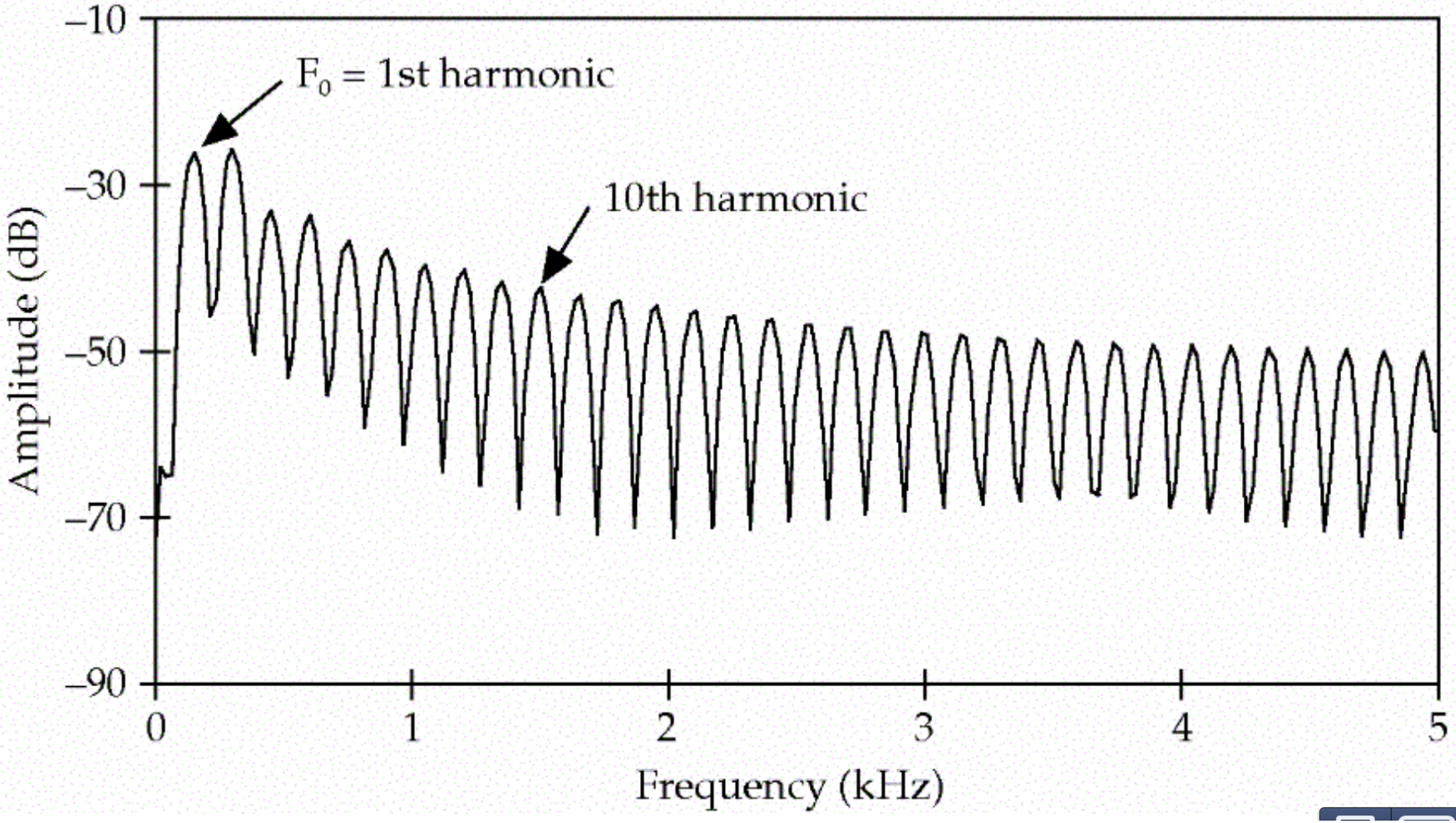
Vocal Fold Cycles



The vocal source at 150 Hz



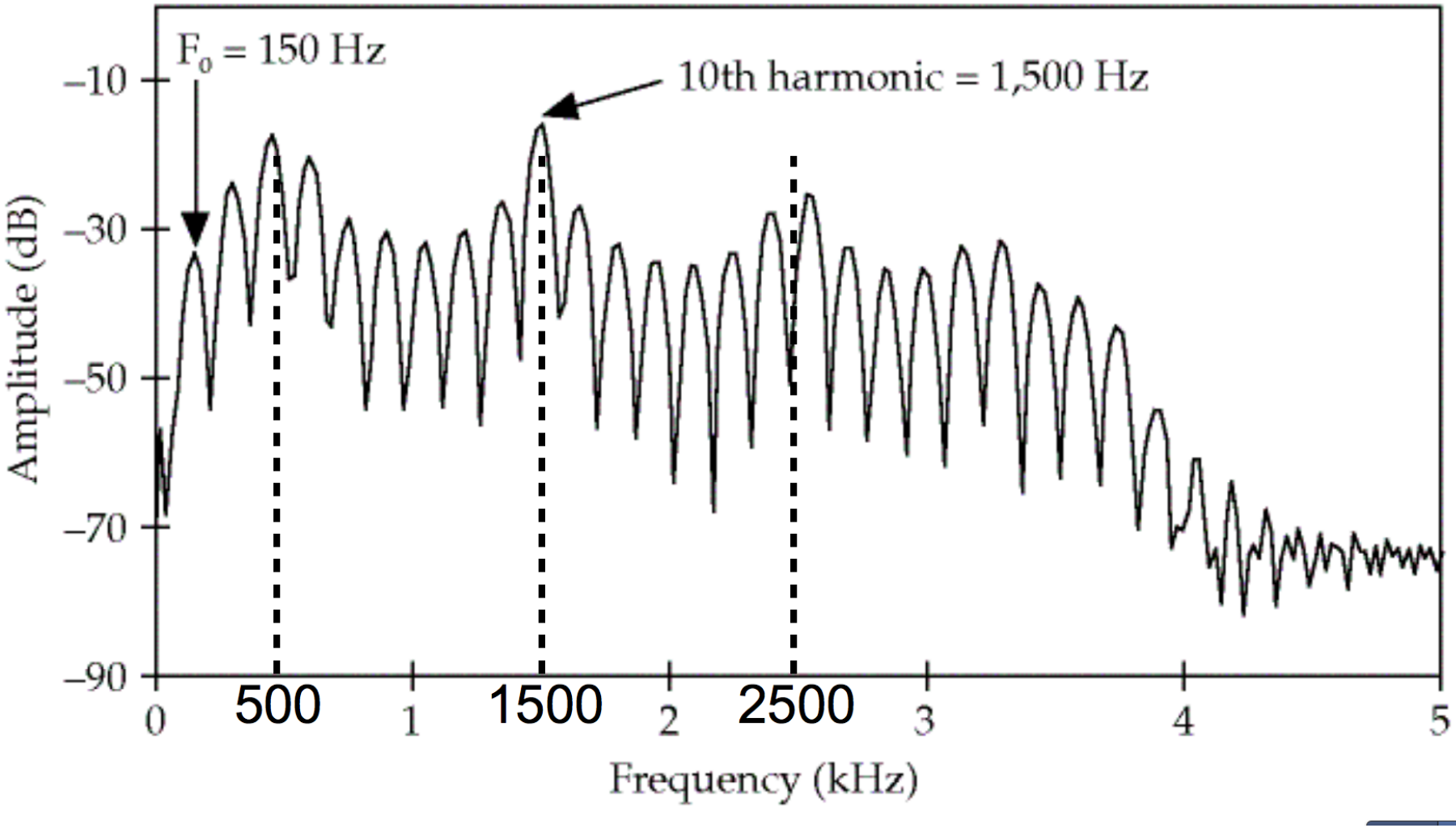
The harmonics



Source filter model of vowels

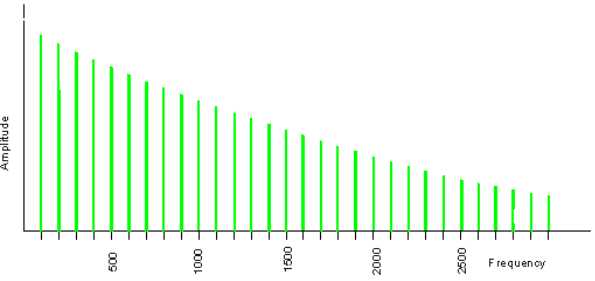
- Any body of air will vibrate in a way that depends on its size and shape.
- Vocal tract as "amplifier"; amplifies certain harmonics
- Formants are result of different shapes of vocal tract.

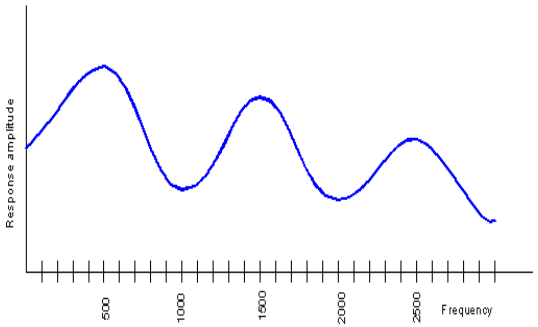
The oral cavity amplifies some harmonics

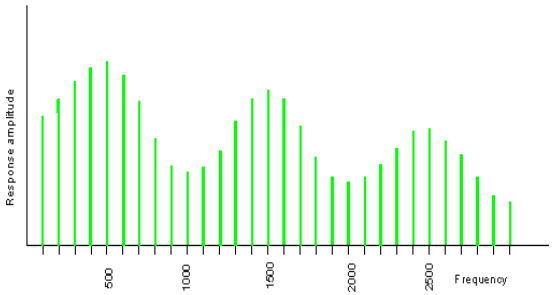


Source-filter model of speech production

Input → Filter → Output







Glottal spectrum

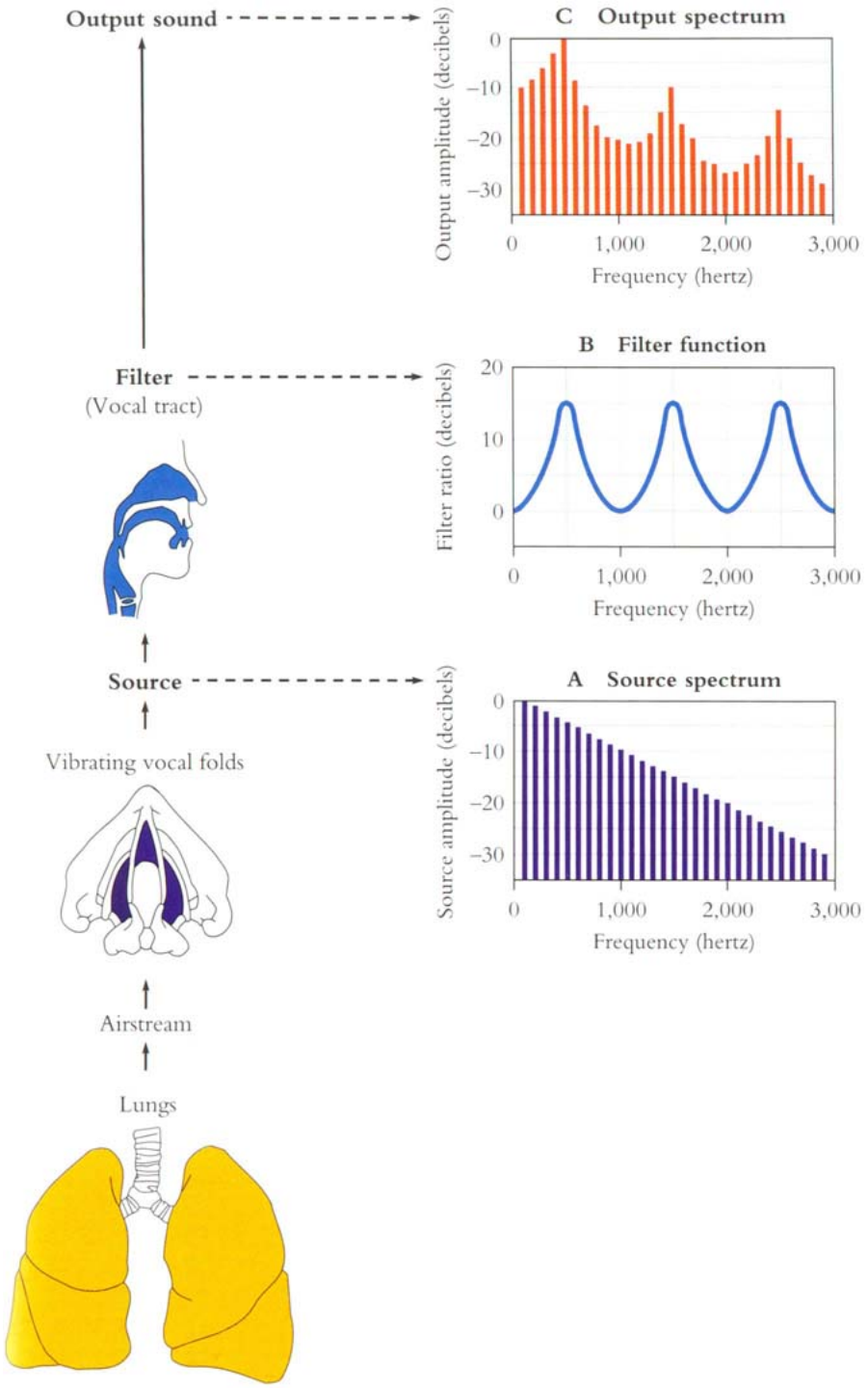
Vocal tract frequency response function

Source and filter are independent, so:

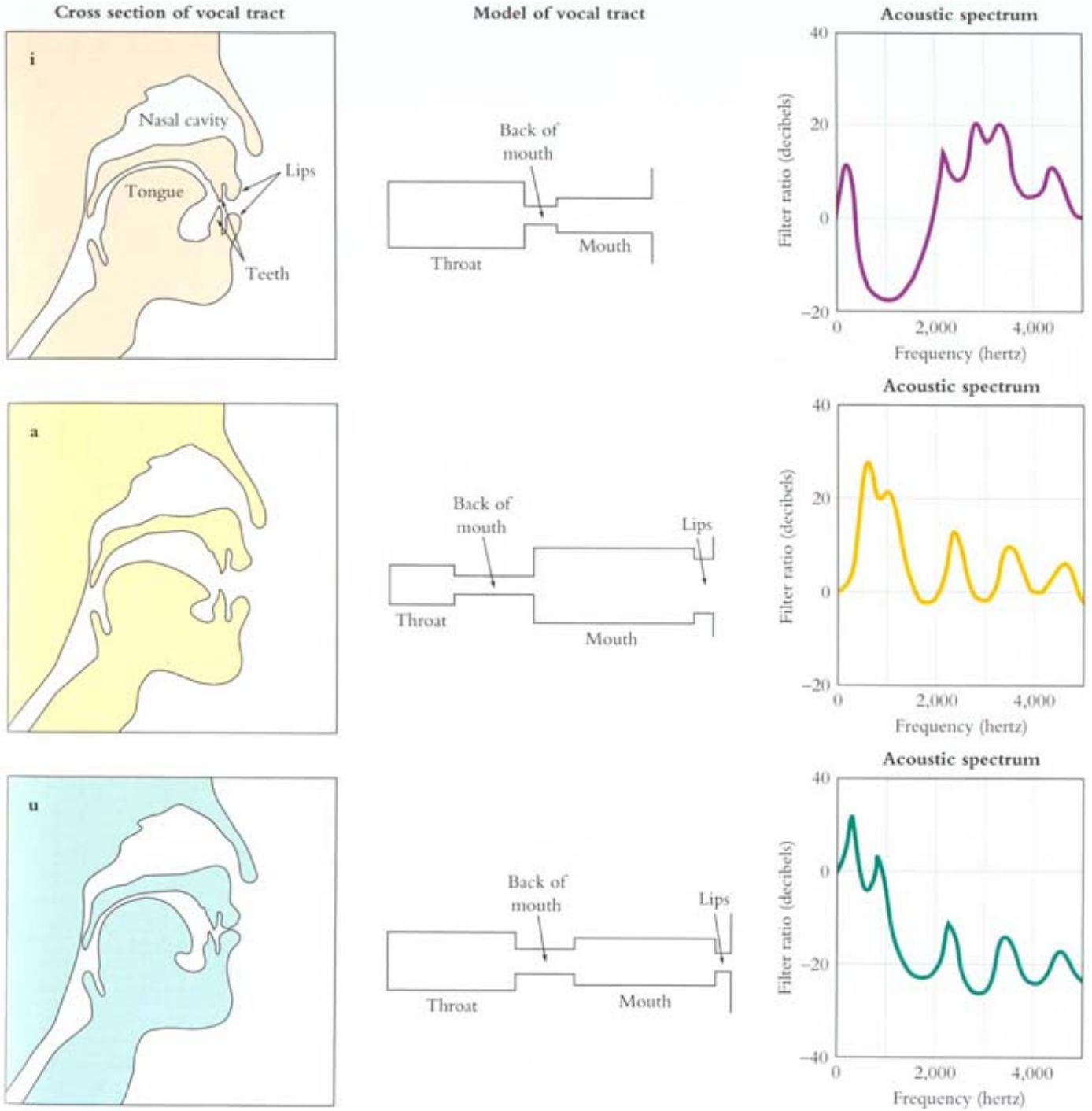
Different vowels can have same pitch

The same vowel can have different pitch

Figures and text from Ratree Wayland slide from his website


Output sound
C Output spectrum
Output amplitude (decibels)
0
−10
−20
−30
0
1,000
2,000
3,000
Frequency (hertz)
Filter
(Vocal tract)
B Filter function
Filter ratio (decibels)
20
10
0
0
1,000
2,000
3,000
Frequency (hertz)
Source
A Source spectrum
Source amplitude (decibels)
0
−10
−20
−30
0
1,000
2,000
3,000
Frequency (hertz)
Vibrating vocal folds
Airstream
Lungs

From Mark Liberman's Web site



Resonances of the vocal tract

- The human vocal tract as an open tube

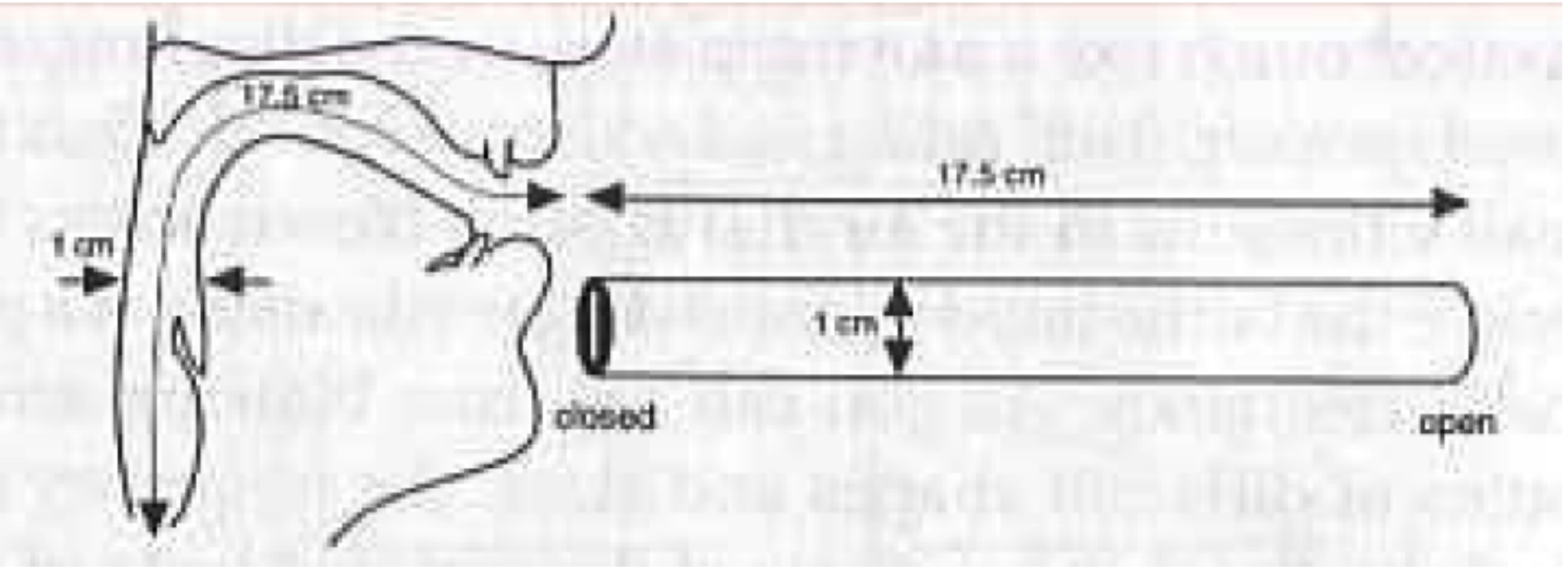
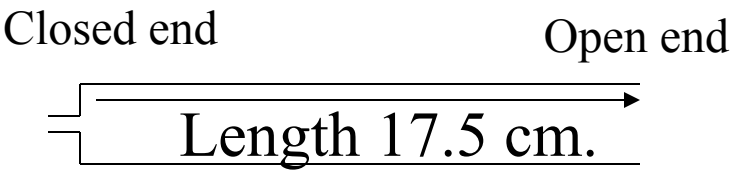


Figure from Ladefoged(1996) p 117

Resonances of the vocal tract

- The human vocal tract as an open tube



- Air in a tube of a given length will tend to vibrate at resonance frequency of tube.

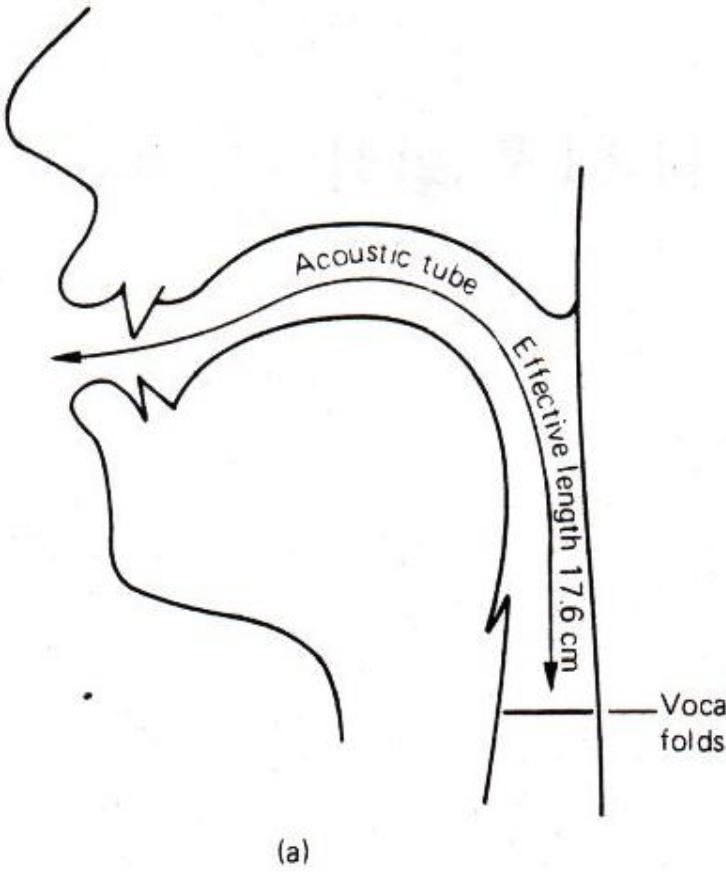
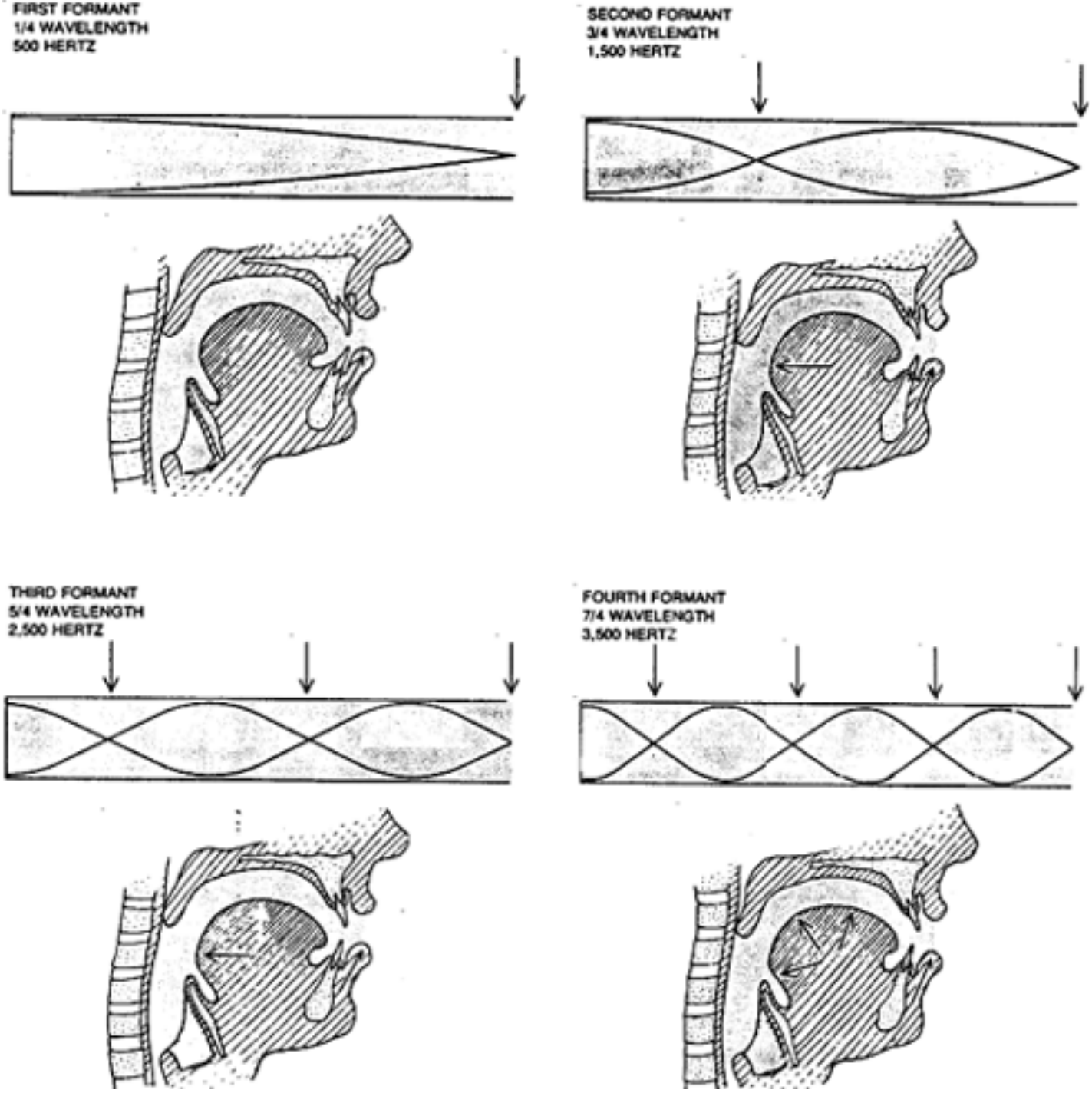


Figure from W. Barry Speech Science slides

Resonances of the vocal tract

- If vocal tract is cylindrical tube open at one end
- Standing waves form in tubes
- Waves will resonate if their wavelength corresponds to dimensions of tube
- Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end.
- Next slide shows what kind of length of waves can fit into a tube with this contraint



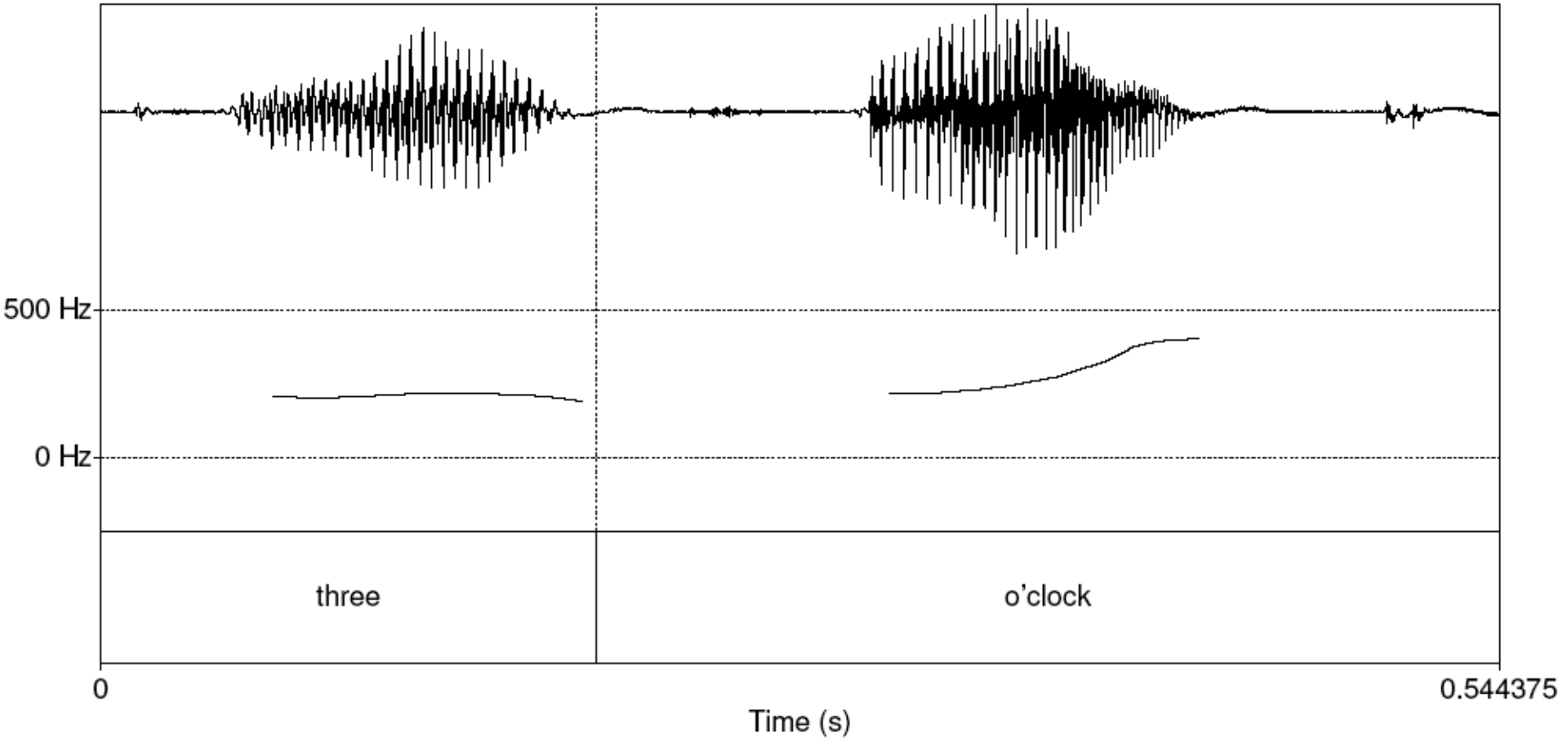
From Sundberg

Defining Intonation

- Ladd (1996) "Intonational phonology"
- "The use of suprasegmental phonetic features

  Suprasegmental = above & beyond the segment/phone
  - F0
  - 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.

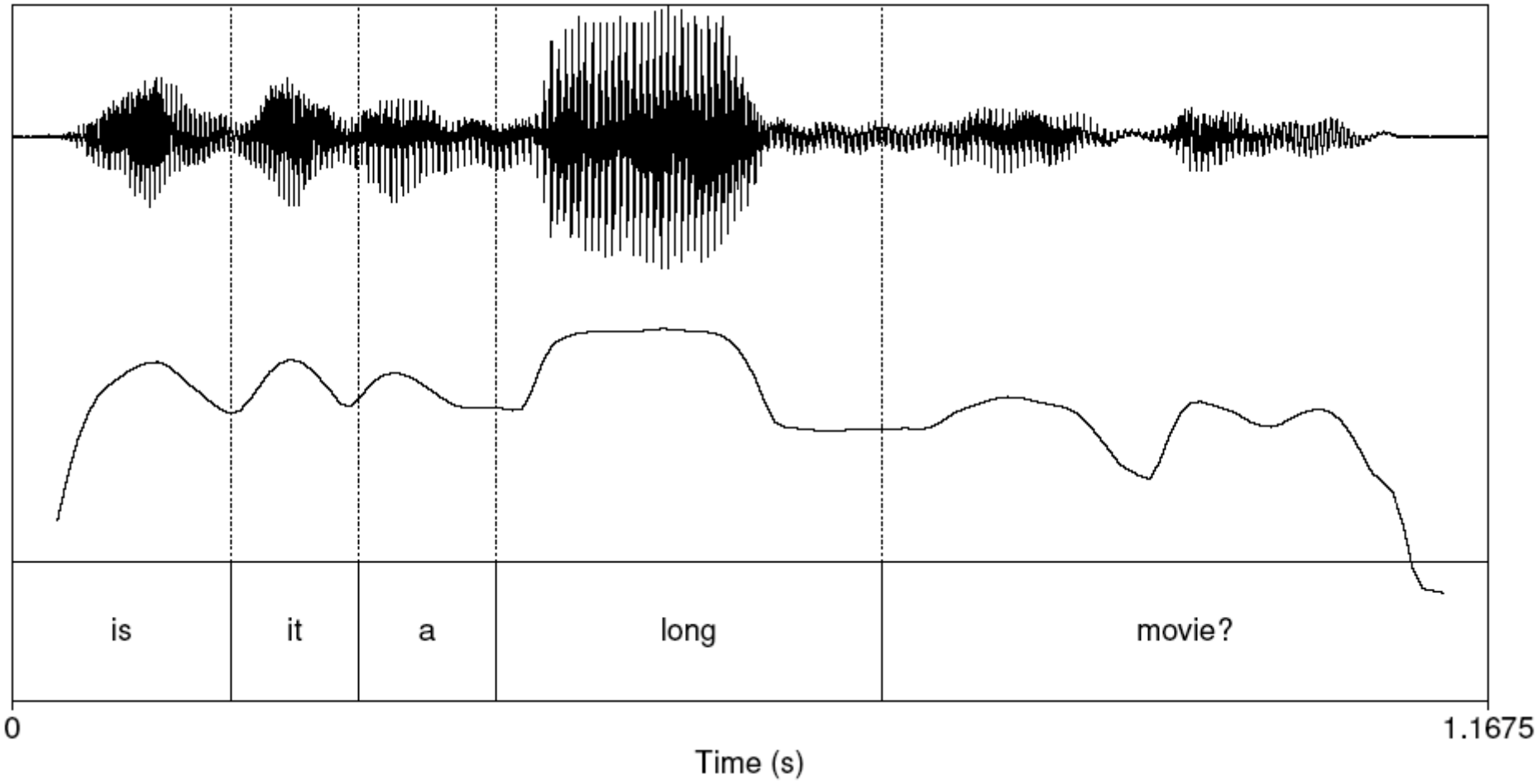
Pitch track



Pitch is not Frequency

- Pitch is the mental sensation or perceptual correlated 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.

From Ladd (1996)

Prosodic Boundaries

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

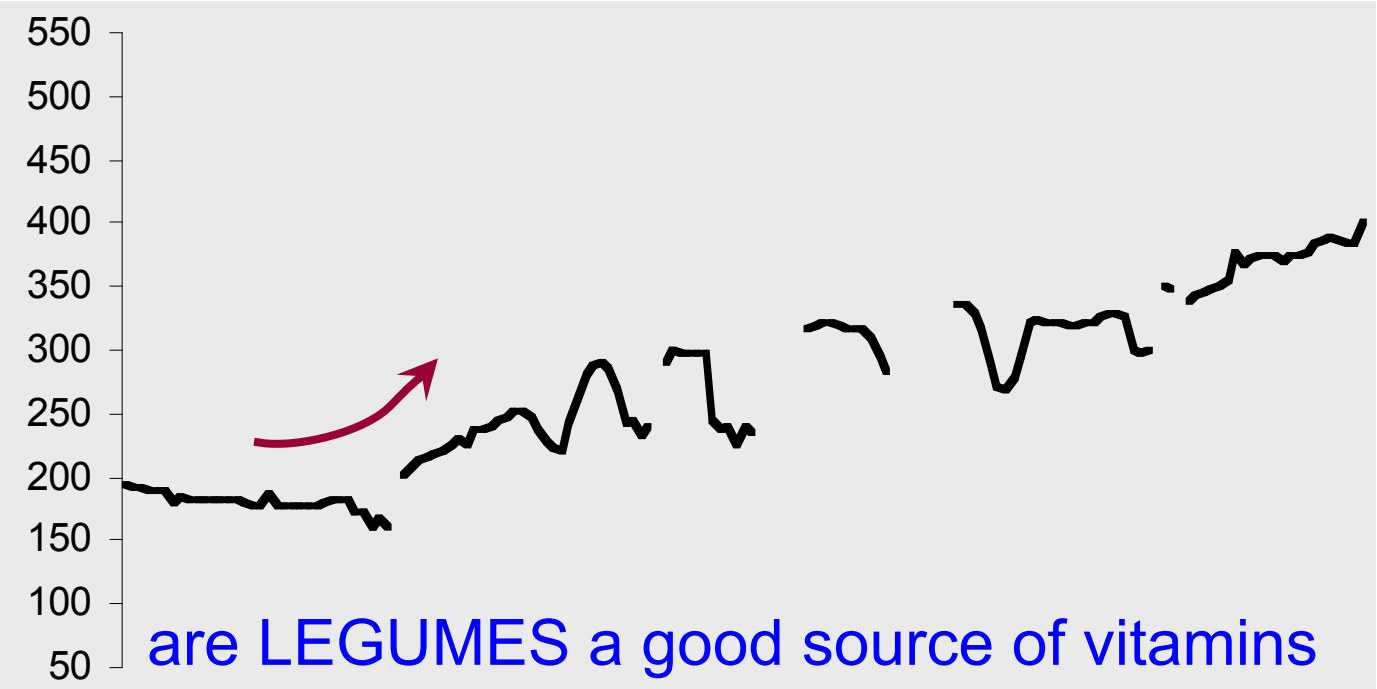
French [bread and cheese]

[French bread] and [cheese]

Slide from Jennifer Venditti

Intonational tunes

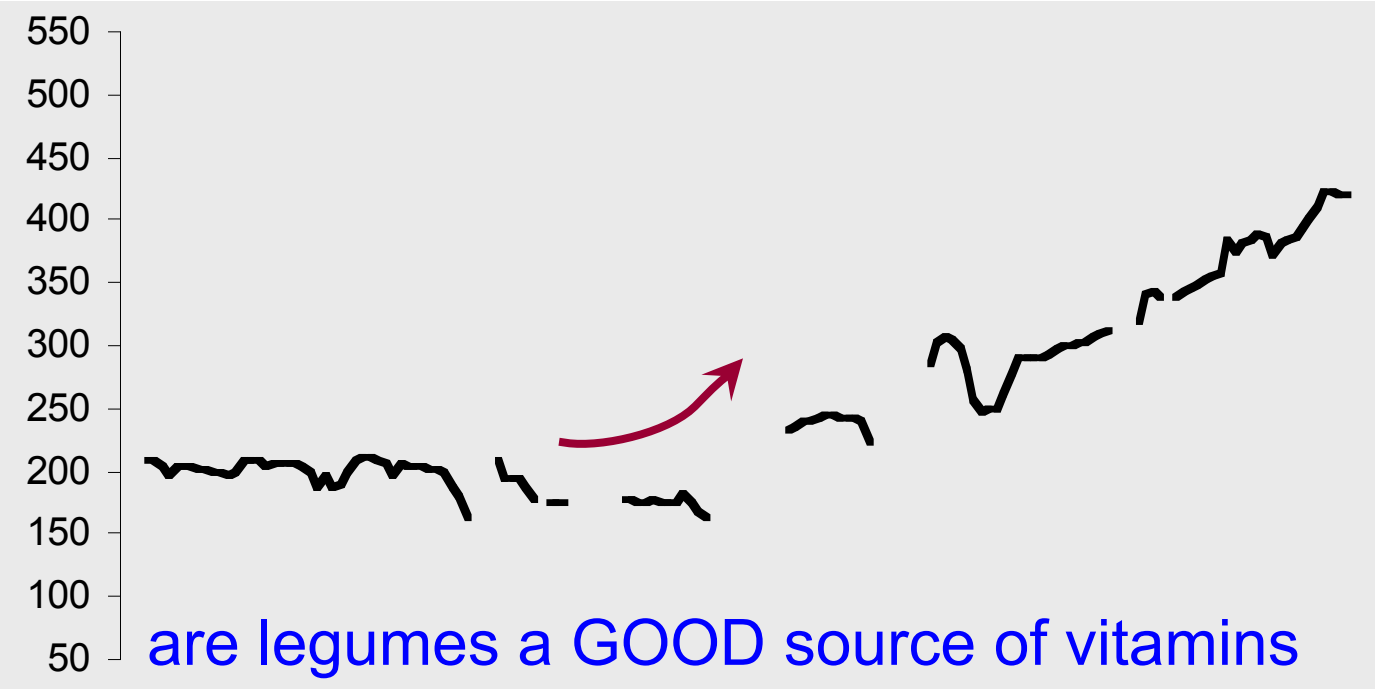
Yes-No question tune



Rise from the main accent to the end of the sentence.

Slide from Jennifer Venditti

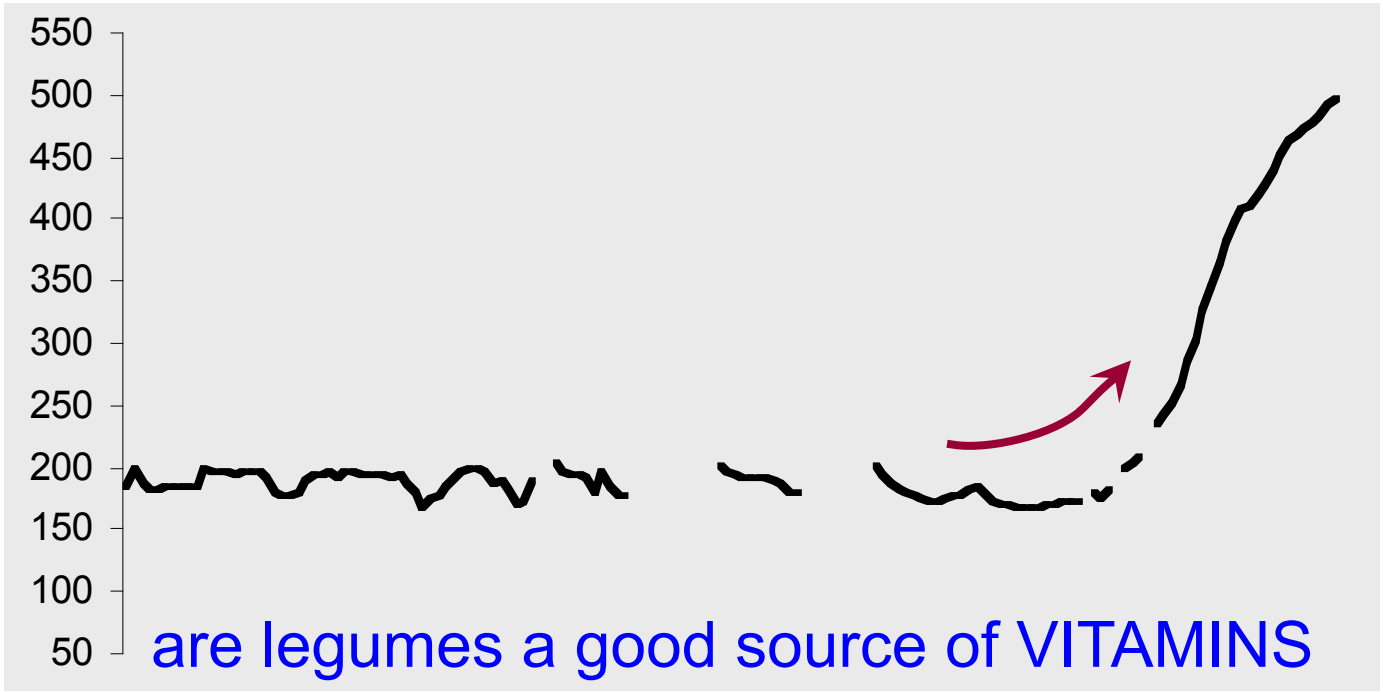
Yes-No question tune



Rise from the main accent to the end of the sentence.

Slide from Jennifer Venditti

Yes-No question tune

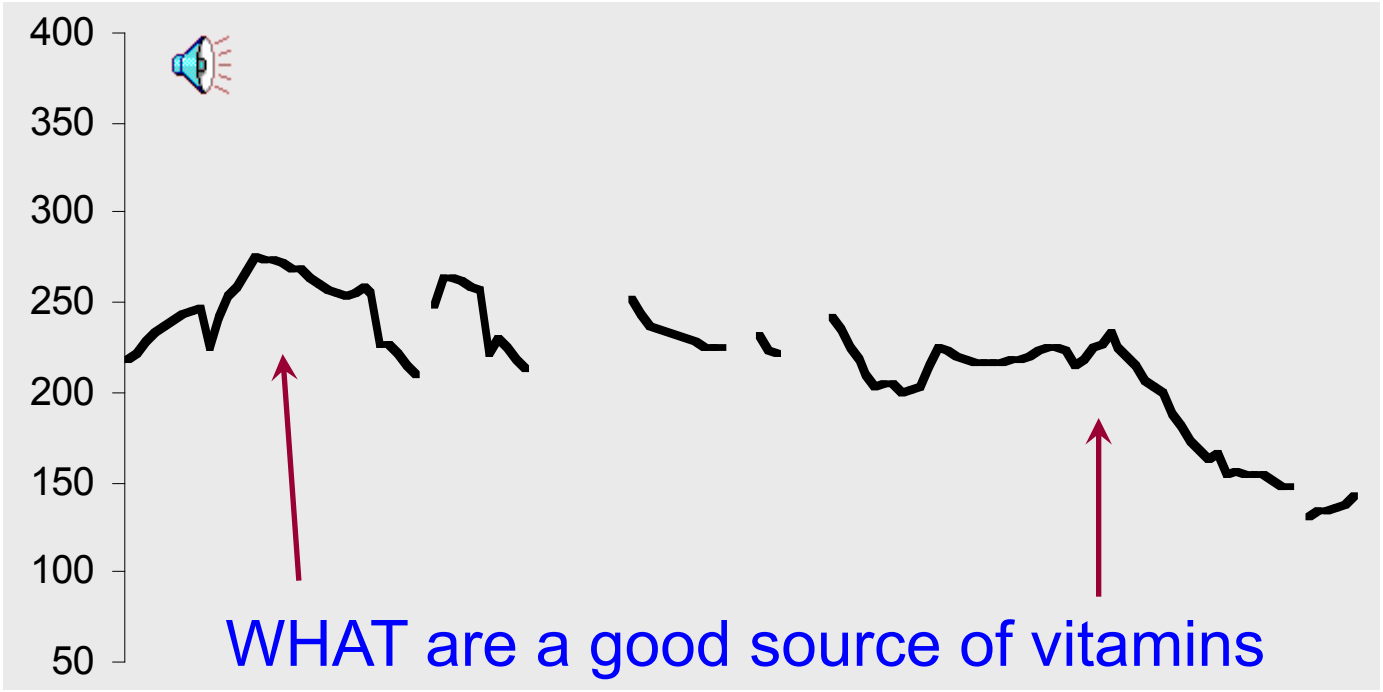


Rise from the main accent to the end of the sentence.

Slide from Jennifer Venditti

WH-questions

[I know that many natural foods are healthy, but ...]

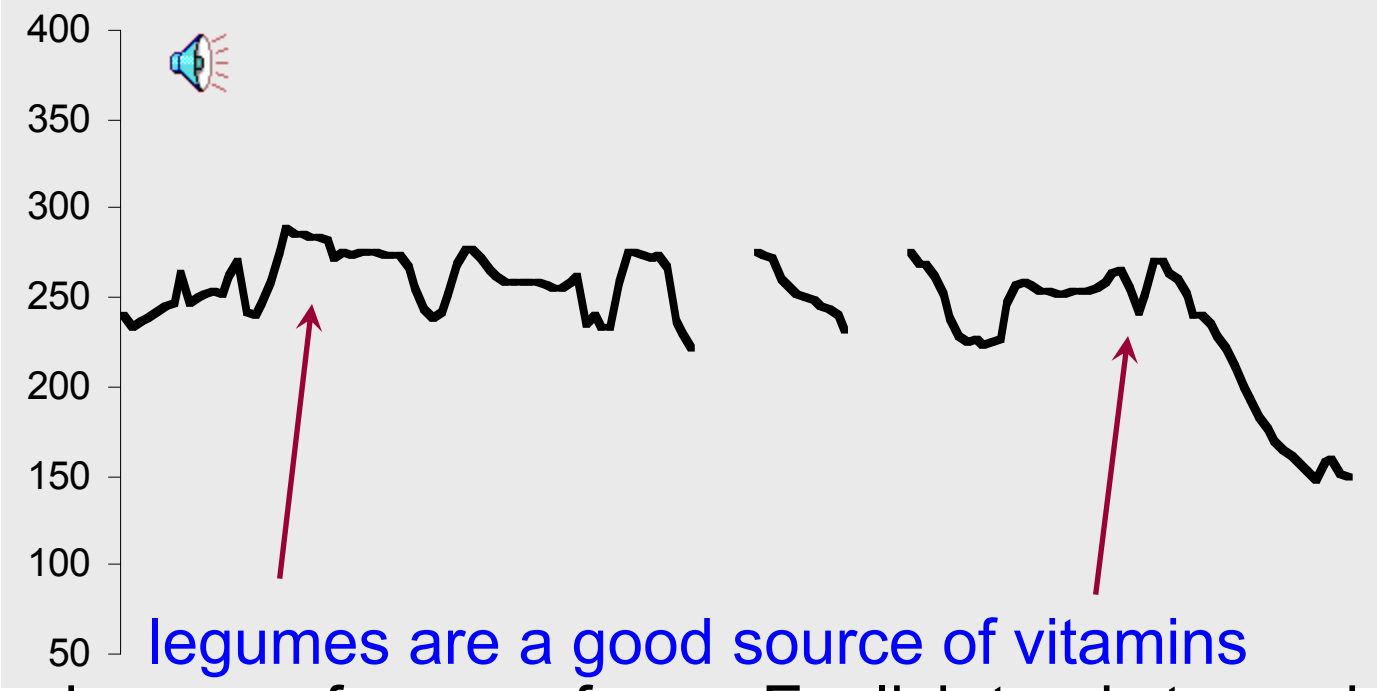


WH-questions typically have falling contours, like statements.

Slide from Jennifer Venditti

Broad focus

“Tell me something about the world.”

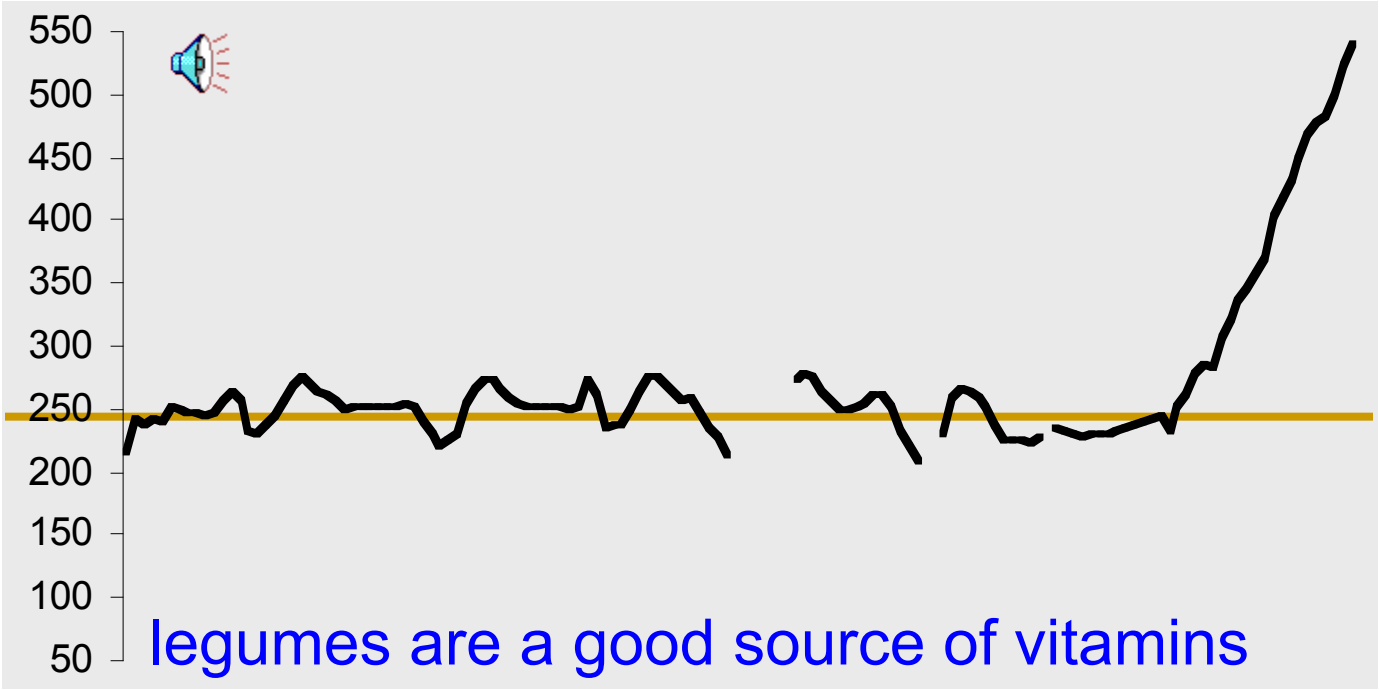


In the absence of narrow focus, English tends to mark the first and last ‘content’ words with perceptually prominent accents.

Slide from Jennifer Venditti

Rising statements

"Tell me something I didn't already know."



[... does this statement qualify?]

High-rising statements can signal that the speaker is seeking approval.

Slide from Jennifer Venditti

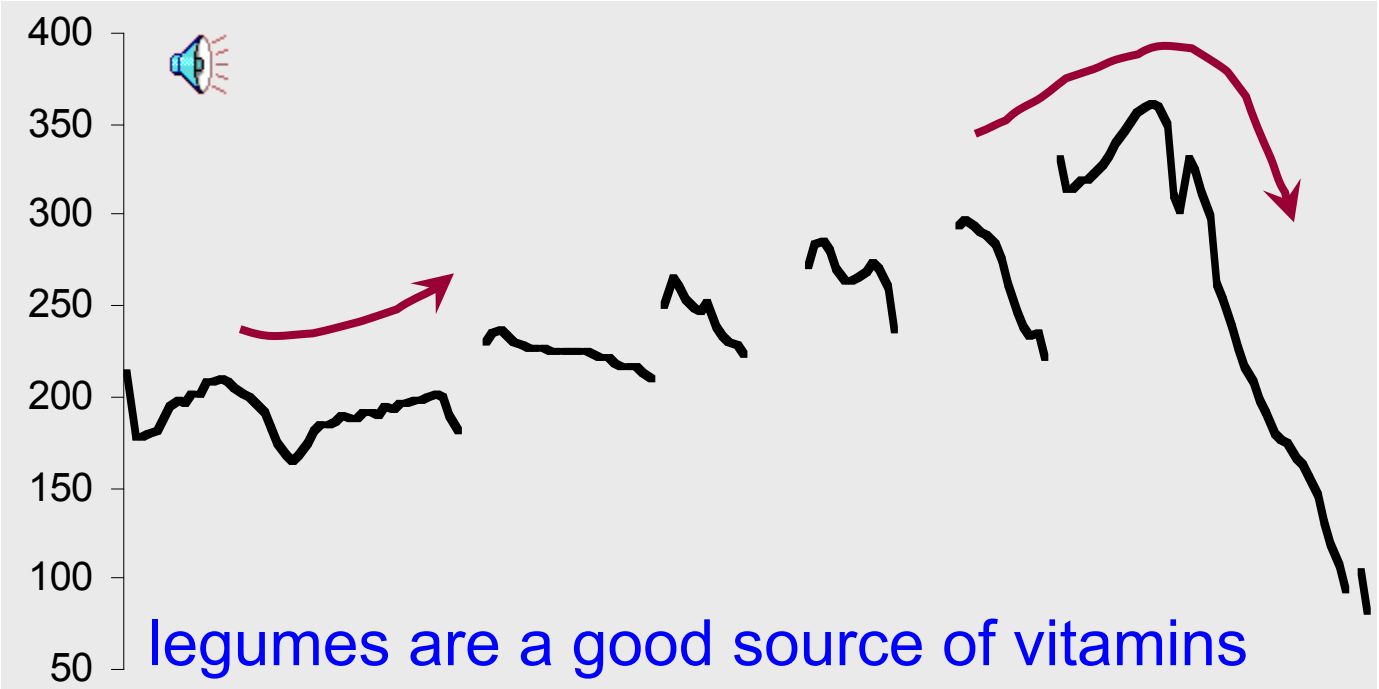
Yes-No question

Rise from the main accent to the end of the sentence.

Slide from Jennifer Venditti

‘Surprise-redundancy’ tune

[How many times do I have to tell you ...]

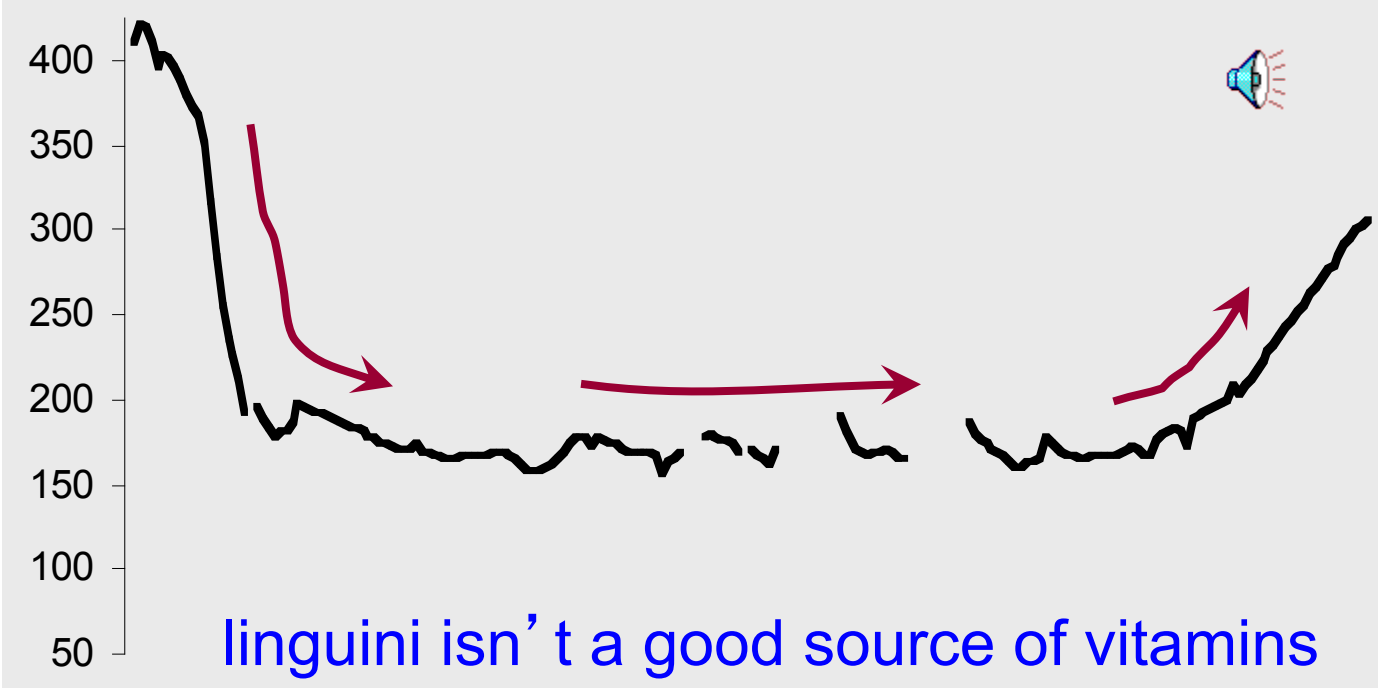


Low beginning followed by a gradual rise to a high at the end.

Slide from Jennifer Venditti

‘Contradiction’ tune

“I’ve heard that linguini is a good source of vitamins.”



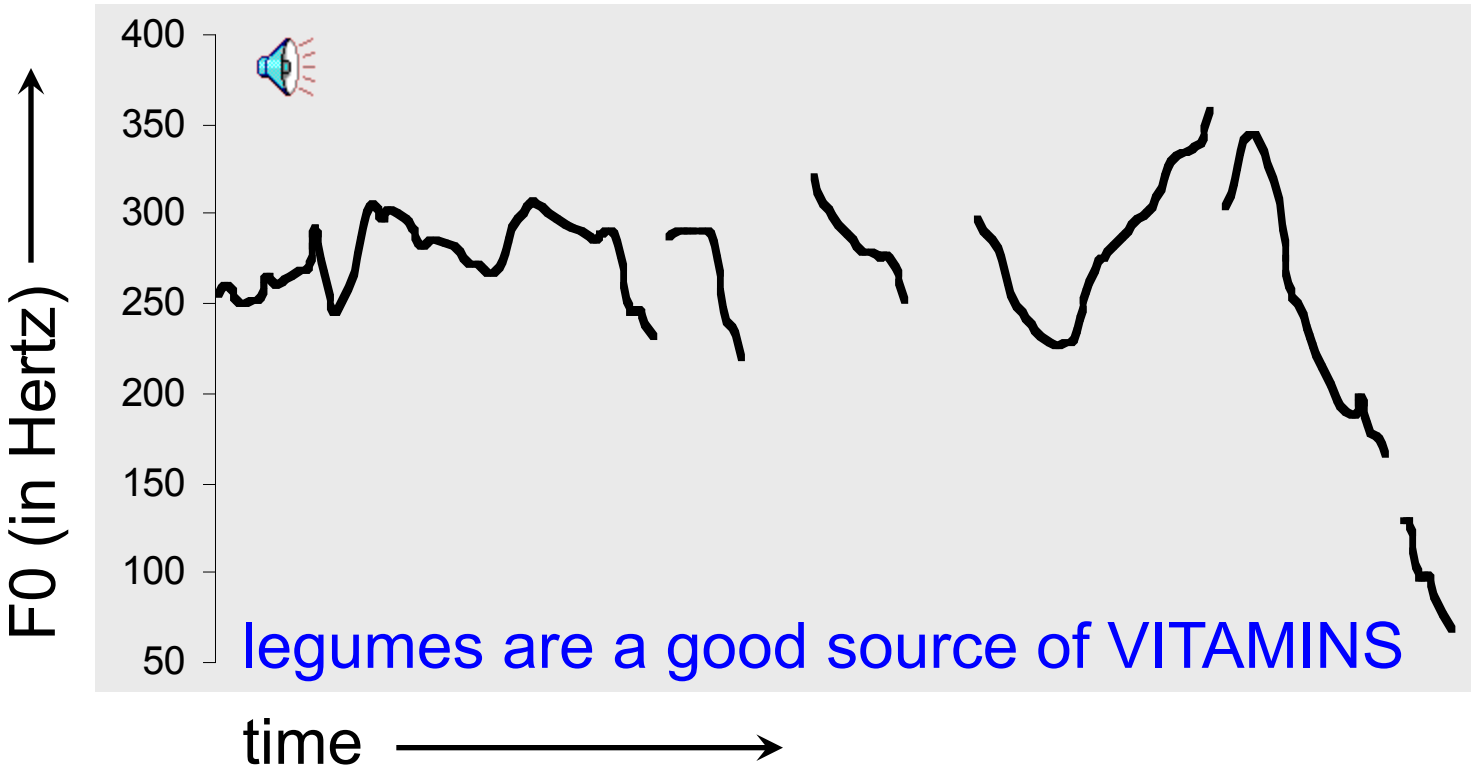
[... how could you think that?]

Sharp fall at the beginning, flat and low, then rising at the end.

Slide from Jennifer Venditti

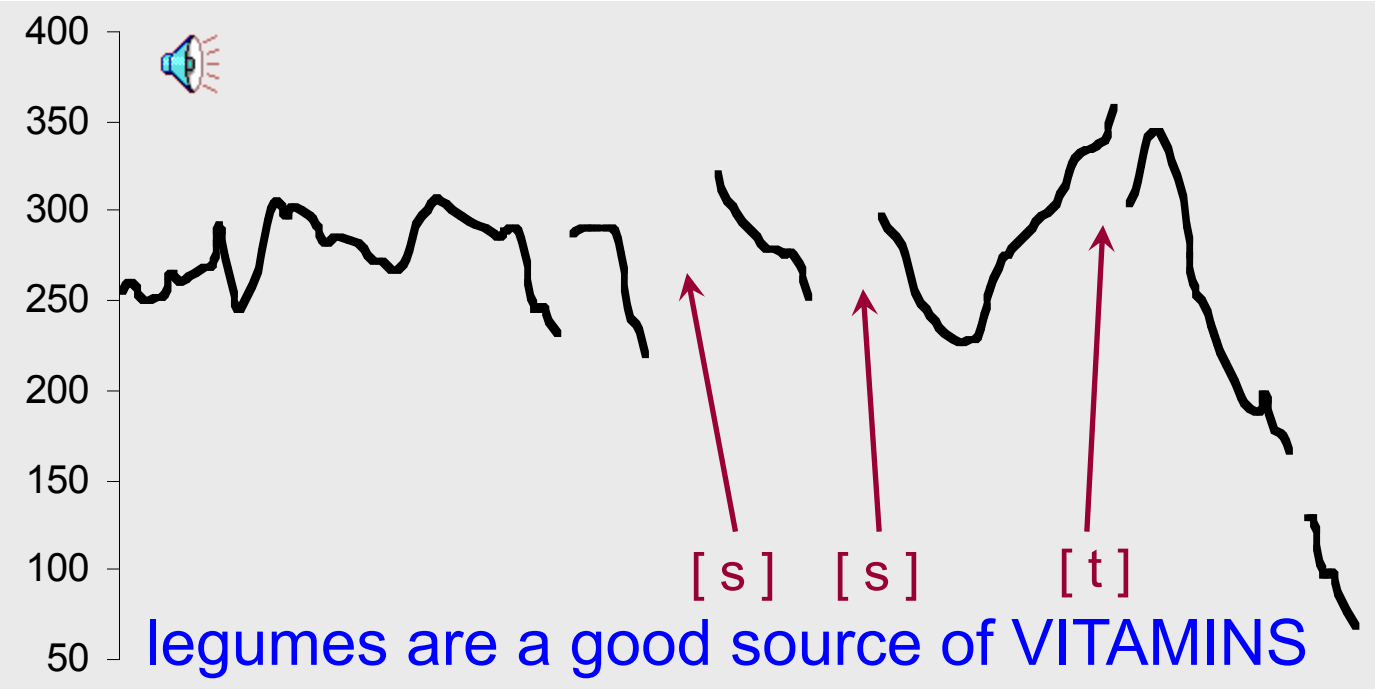
Thinking about F0

Graphic representation of F0



Slide from Jennifer Venditti

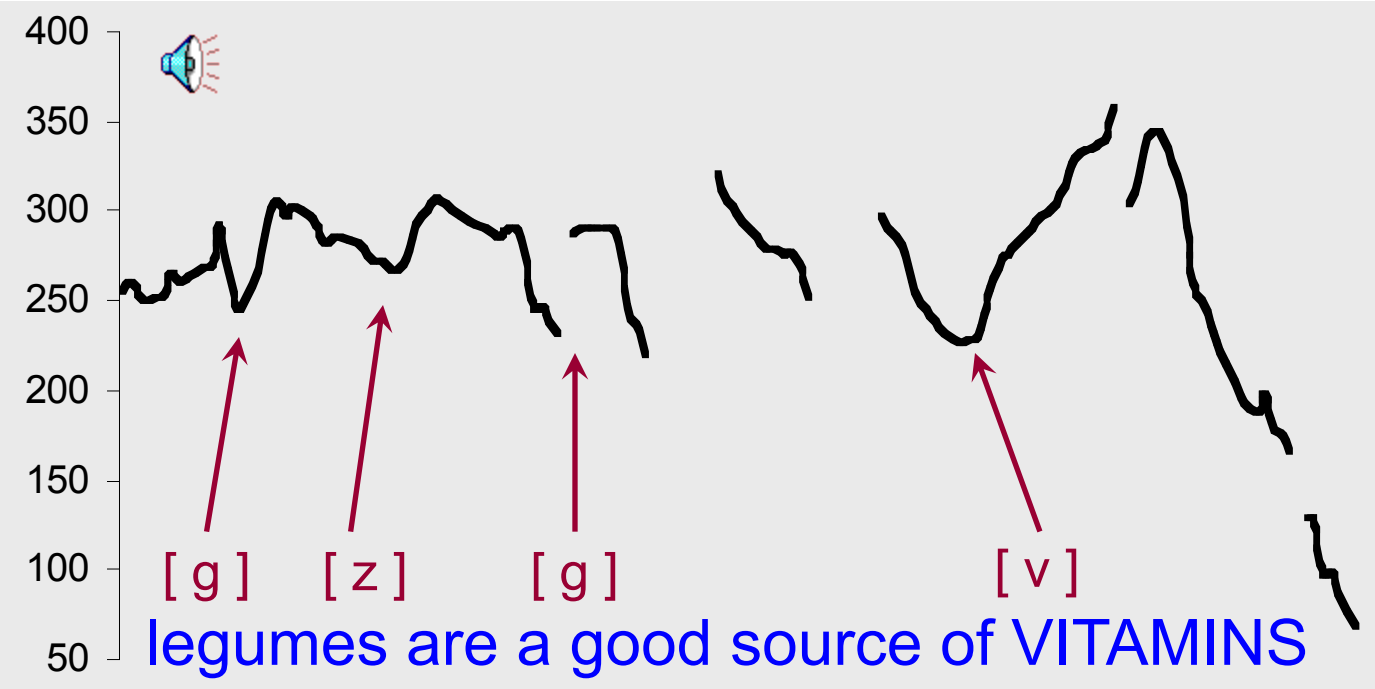
The ‘ripples’



F0 is not defined for consonants without vocal fold vibration.

Slide from Jennifer Venditti

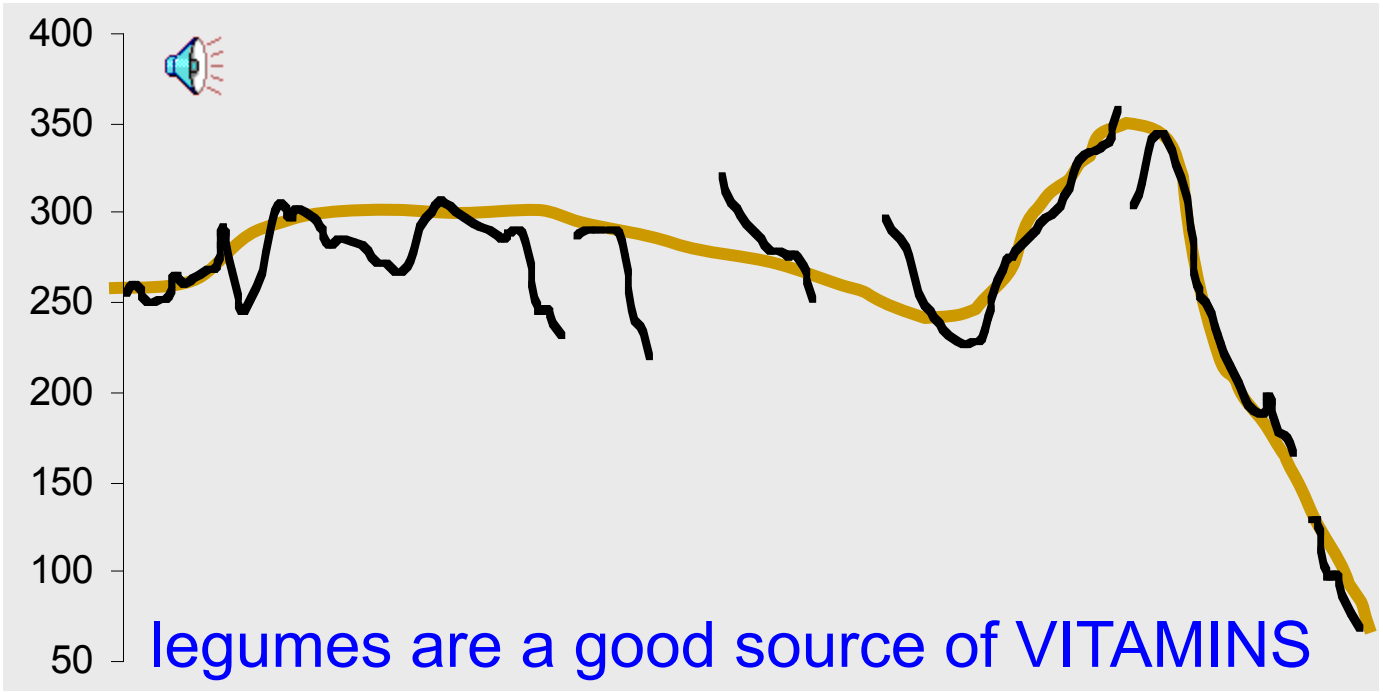
The ‘ripples’



... and F0 can be perturbed by consonants with an extreme constriction in the vocal tract.

Slide from Jennifer Venditti

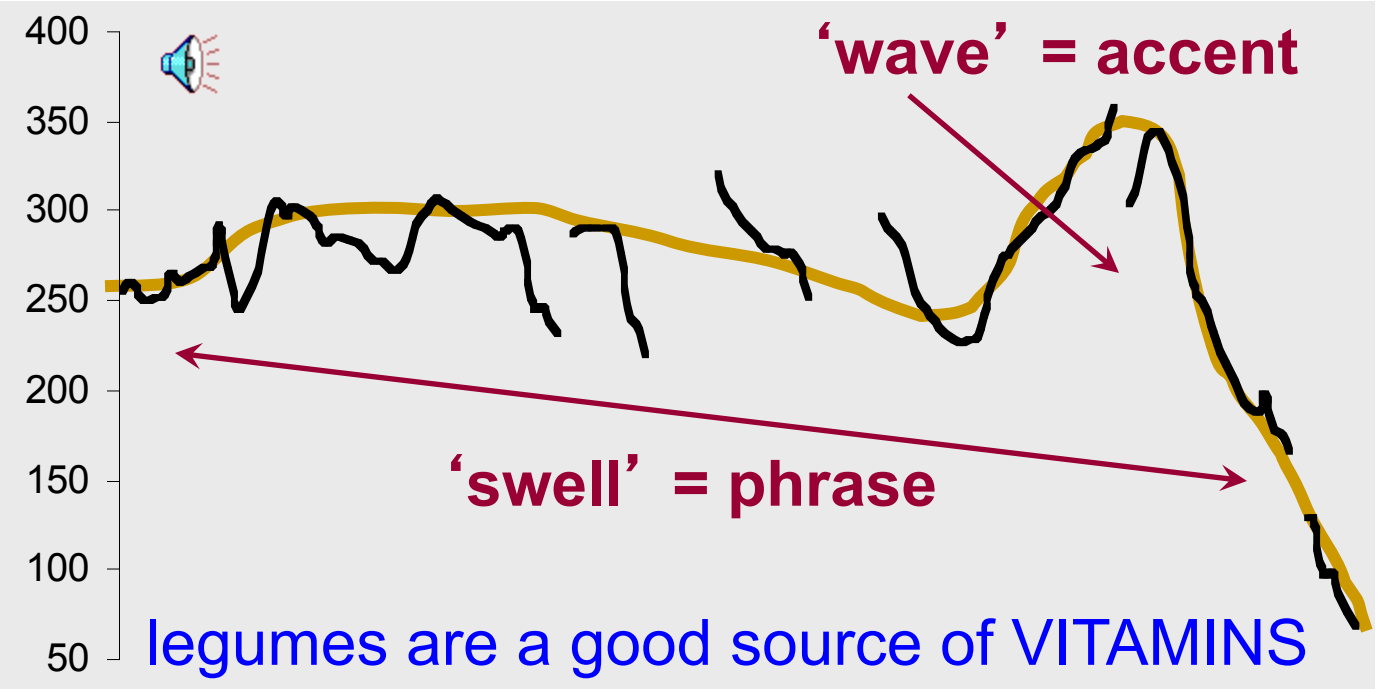
Abstraction of the F0 contour



Our perception of the intonation contour abstracts away from these perturbations.

Slide from Jennifer Venditti

The ‘waves’ and the ‘swells’



Slide from Jennifer Venditti

Prominence:
Placement of Pitch Accents

Stress vs. accent

- *Stress* is a structural property of a word
  - it marks a potential (arbitrary) location for an accent to occur, if there is one.
- *Accent* is a property of a word in <u>context</u>
  - it is a way to mark intonational prominence in order to 'highlight' important words in the discourse.

(x)							(x)		(accented syll)
x							x		stressed syll
x				x			x		full vowels
x	x	x		x	x		x	x	syllables
vi	ta	mins		Ca	li		for	nia	

Slide from Jennifer Venditti

Stress vs. accent (2)

- The speaker decides to make the word vitamin more prominent by accenting it.
- Lexical stress tell us that this prominence will appear on the first syllable, hence VItamin.

- So prosodic prominence is a function of
  - lexicon
  - context

- I'm a little **surPRISED** to hear it **CHARacterized** as **upBEAT**

Which word receives an accent?

- It depends on the context.
  - The 'new' information in the answer to a question is often accented
    - while the 'old' information is usually not.

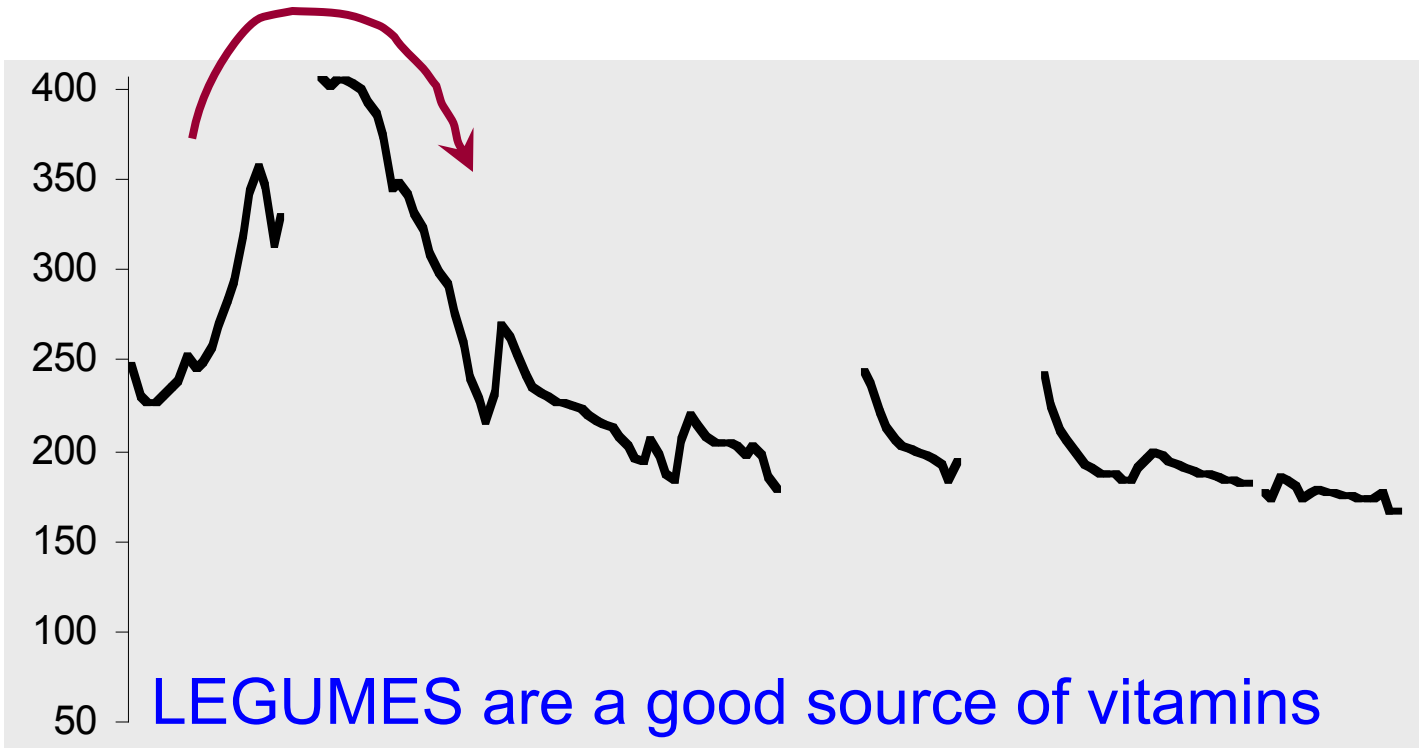
  - Q1: What types of foods are a good source of vitamins?
  - A1: LEGUMES are a good source of vitamins.

  - Q2: Are legumes a source of vitamins?
  - A2: Legumes are a GOOD source of vitamins.

  - Q3: I've heard that legumes are healthy, but what are they a good source of ?
  - A3: Legumes are a good source of VITAMINS.

Slide from Jennifer Venditti

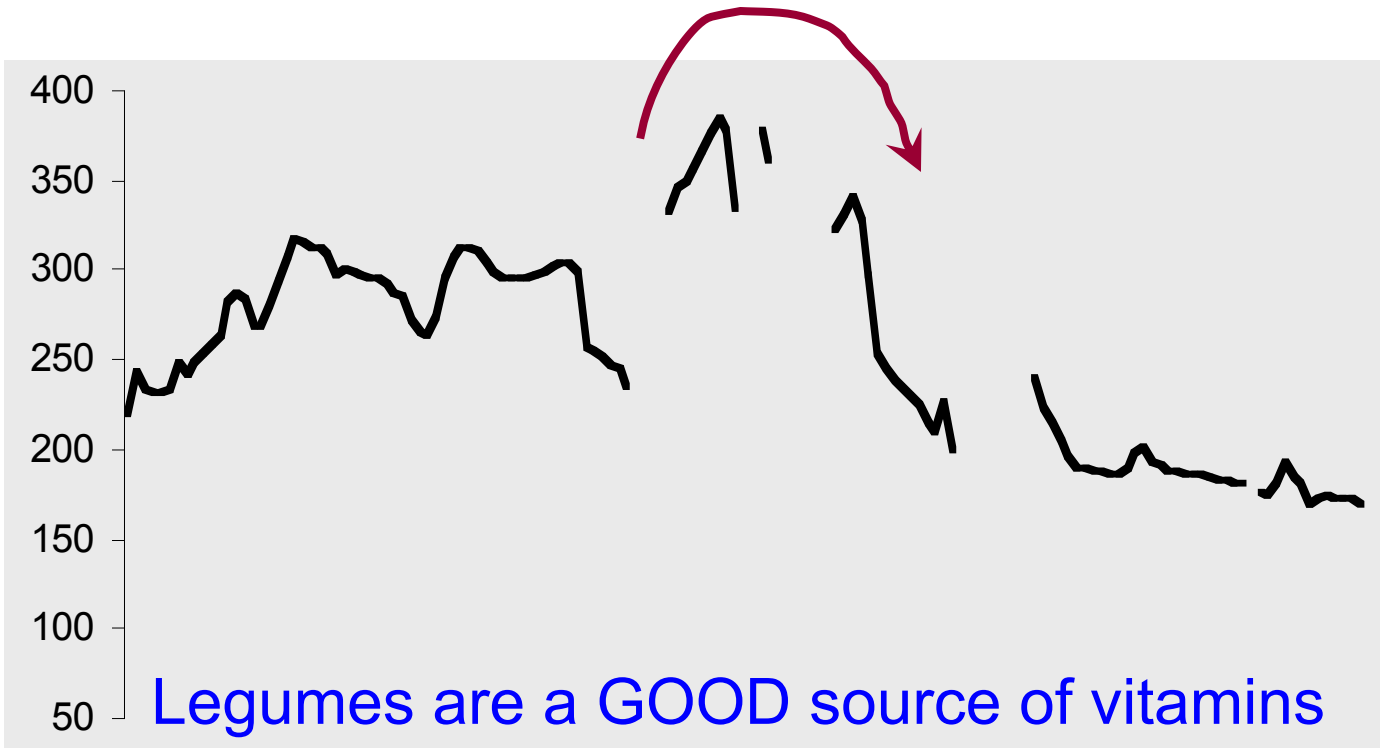
Same ‘tune’, different alignment



The main **rise-fall** accent (= “I assert this”) shifts locations.

Slide from Jennifer Venditti

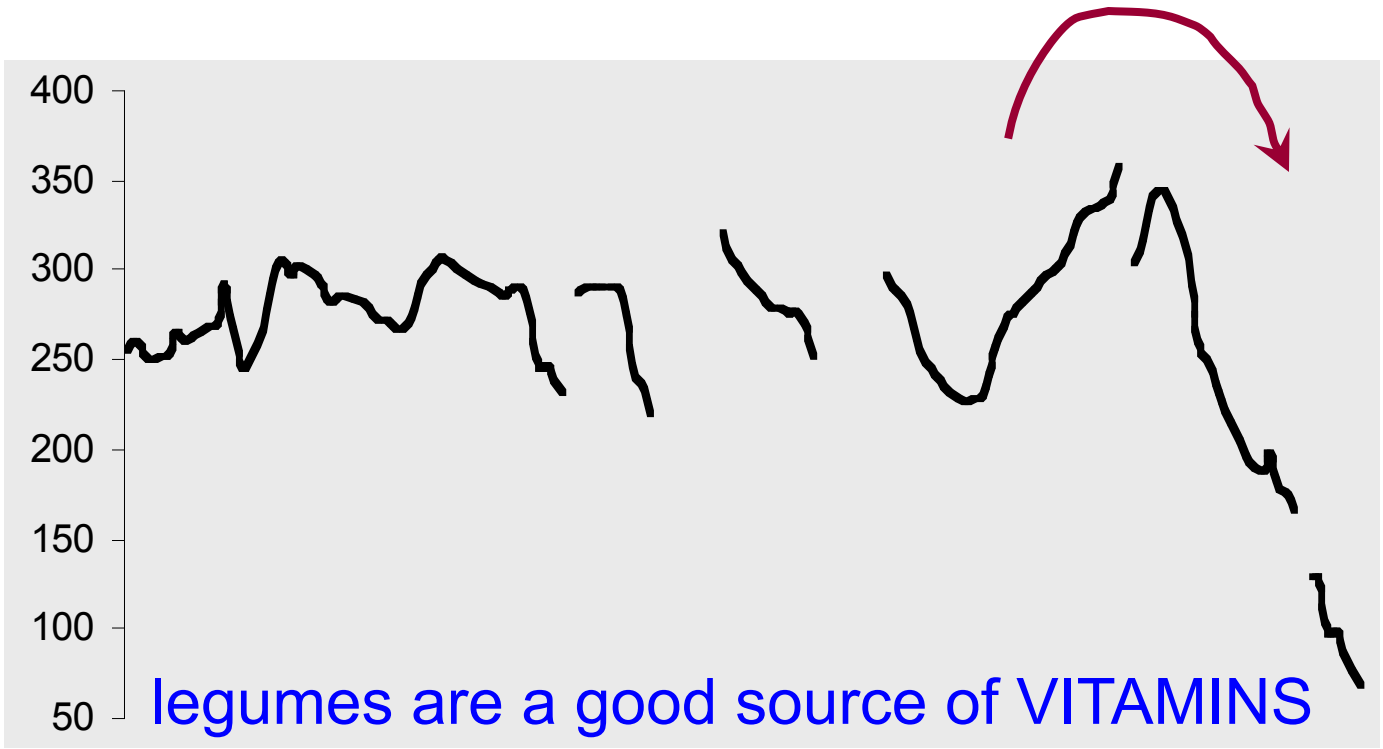
Same ‘tune’, different alignment



The main **rise-fall** accent (= “I assert this”) shifts locations.

Slide from Jennifer Venditti

Same ‘tune’, different alignment



The main **rise-fall** accent (= “I assert this”) shifts locations.

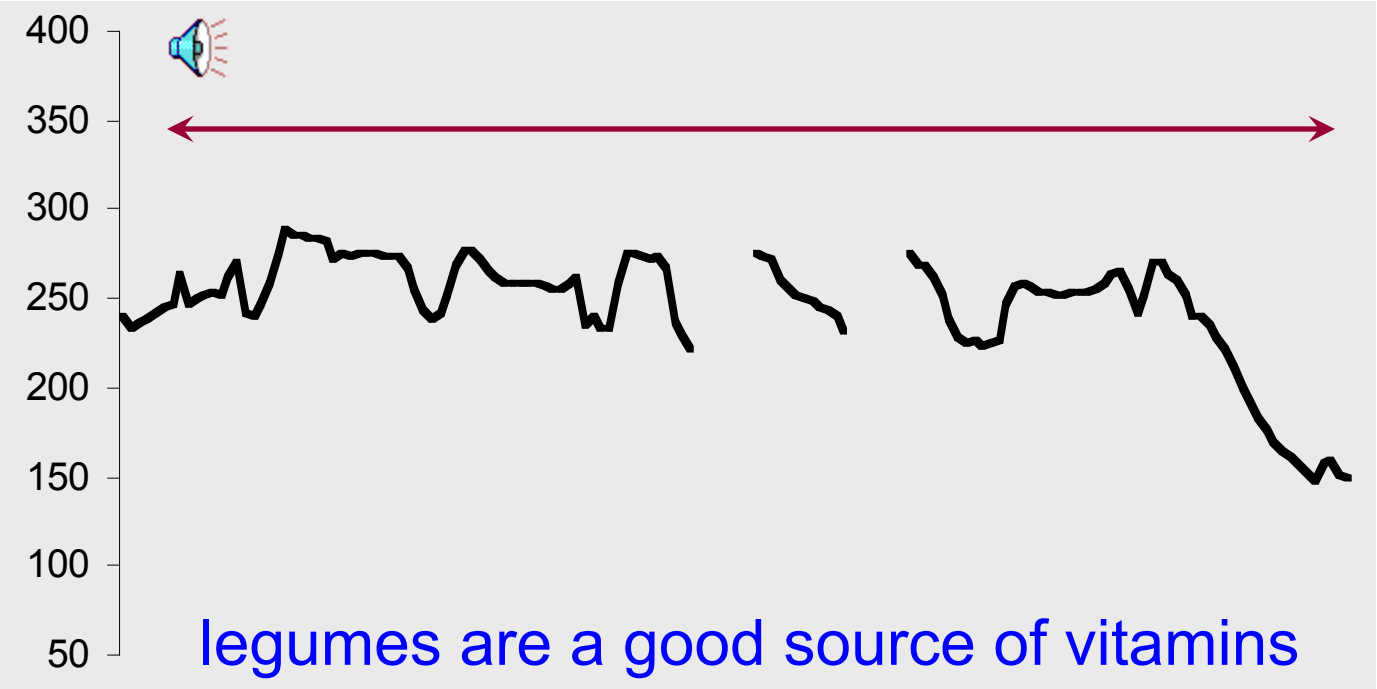
Slide from Jennifer Venditti

Levels of prominence

- Most phrases have more than one accent
- The last accent in a phrase is perceived as more prominent
  - Called the Nuclear Accent
- Emphatic accents like nuclear accent often used for semantic purposes, such as indicating that a word is contrastive, or the semantic focus.
  - The kind of thing you uses ***s in IM, or capitalized letters
  - 'I know **SOMETHING** interesting is sure to happen,' she said to herself.
- Can also have words that are **less** prominent than usual
  - Reduced words, especially function words.
- Often use 4 classes of prominence:
  - **Emphatic accent, pitch accent, unaccented, reduced**

Intonational phrasing/boundaries

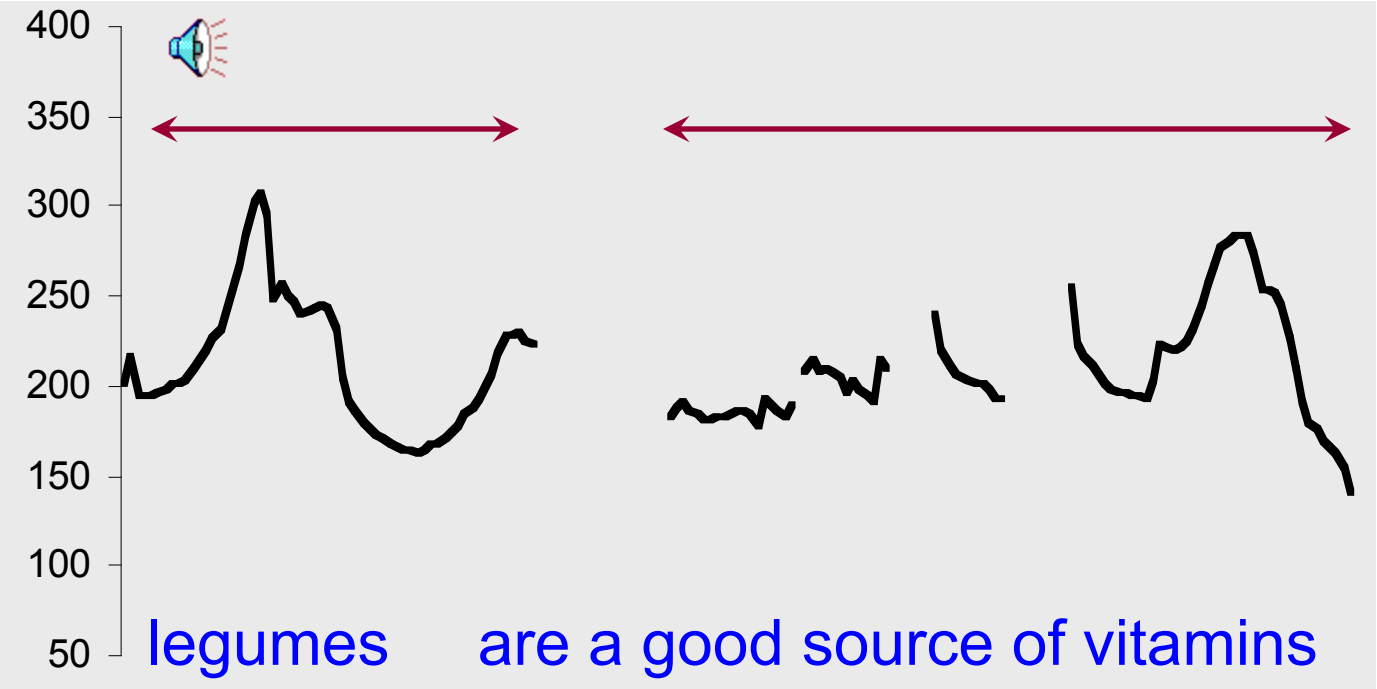
A single intonation phrase



Broad focus statement consisting of one intonation phrase (that is, one intonation tune spans the whole unit).

Slide from Jennifer Venditti

Multiple phrases



Utterances can be ‘chunked’ up into smaller phrases
in order to signal the importance of information in each unit.

Slide from Jennifer Venditti

Phrasing sometimes helps disambiguate

- **Global ambiguity:**

  The old men and women stayed home.

  Sally saw the man with the binoculars.

  John doesn't drink because he's unhappy.

Slide from Jennifer Venditti

Phrasing can disambiguate

- **Global ambiguity:**

  The old men and women stayed home.
  The old men **%** and women **%** stayed home.

  Sally saw **%** the man with the binoculars.
  Sally saw the man **%** with the binoculars.

  John doesn't drink because he's unhappy.
  John doesn't drink **%** because he's unhappy.

Slide from Jennifer Venditti

Phrasing sometimes helps disambiguate

- **Temporary ambiguity:**

  When Madonna sings the song ...

Slide from Jennifer Venditti

Phrasing sometimes helps disambiguate

- **Temporary ambiguity:**

  When Madonna sings the song is a hit.

Slide from Jennifer Venditti

Phrasing sometimes helps disambiguate

- **Temporary ambiguity:**

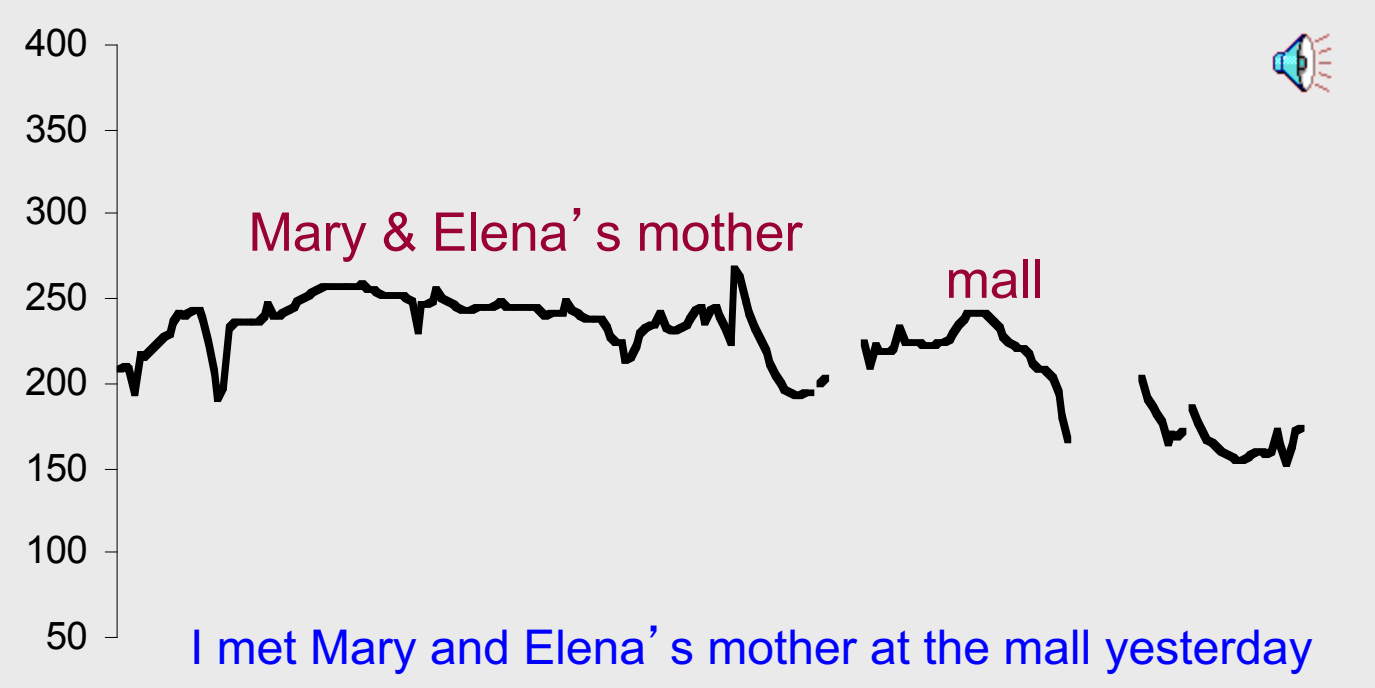
  When Madonna sings **%** the song is a hit.

  When Madonna sings the song **%** it's a hit.

  [from Speer & Kjelgaard (1992)]

Slide from Jennifer Venditti

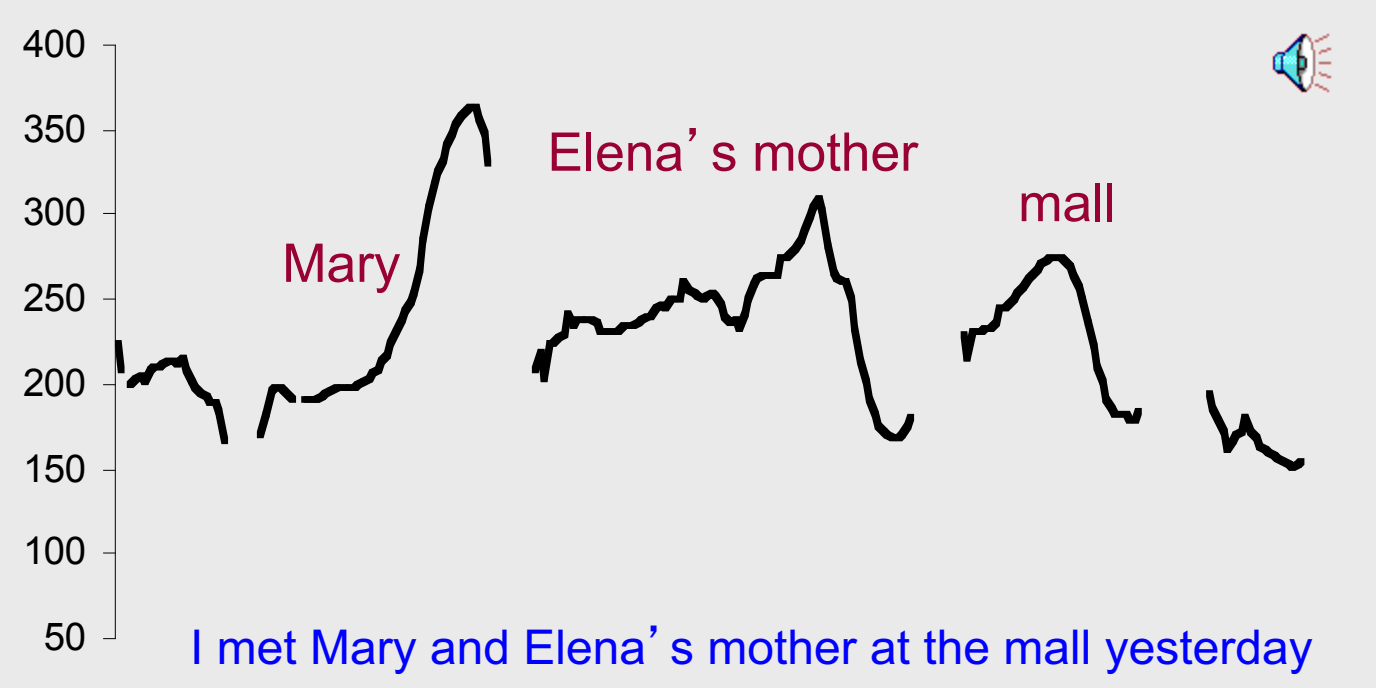
Phrasing sometimes helps disambiguate



One intonation phrase with relatively flat overall pitch range.

Slide from Jennifer Venditti

Phrasing sometimes helps disambiguate



Separate phrases, with expanded pitch movements.

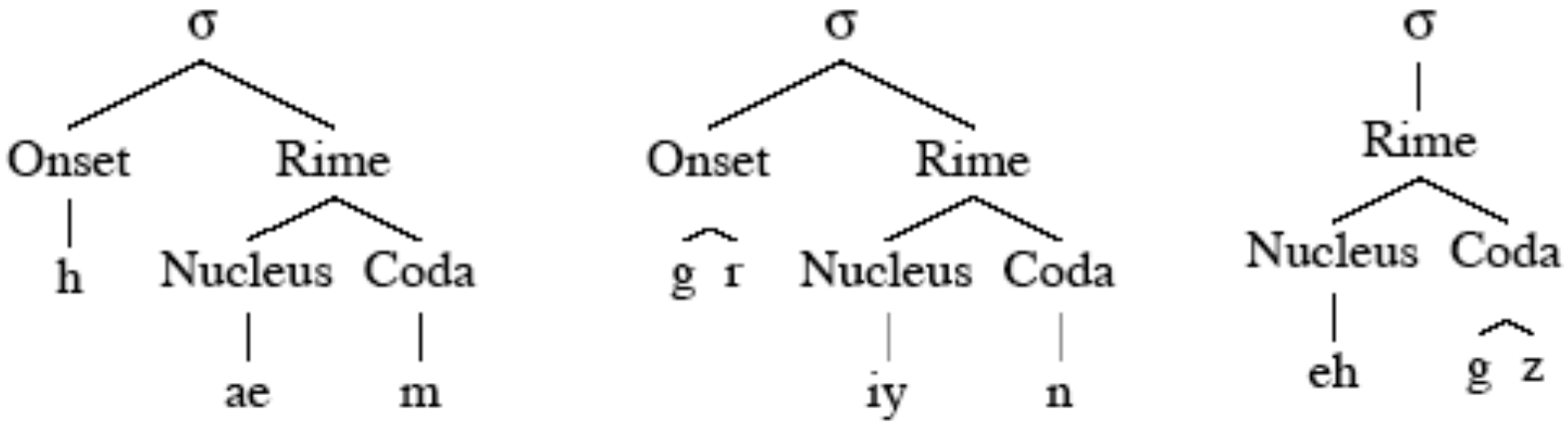
Slide from Jennifer Venditti

Using Intonation in Spoken Language Processing

1) Prominence/Accent: Tells us about focus of utterance
2) Tune: whether utterance is question/statement, important for affect extraction
3) Boundaries: can help parsing

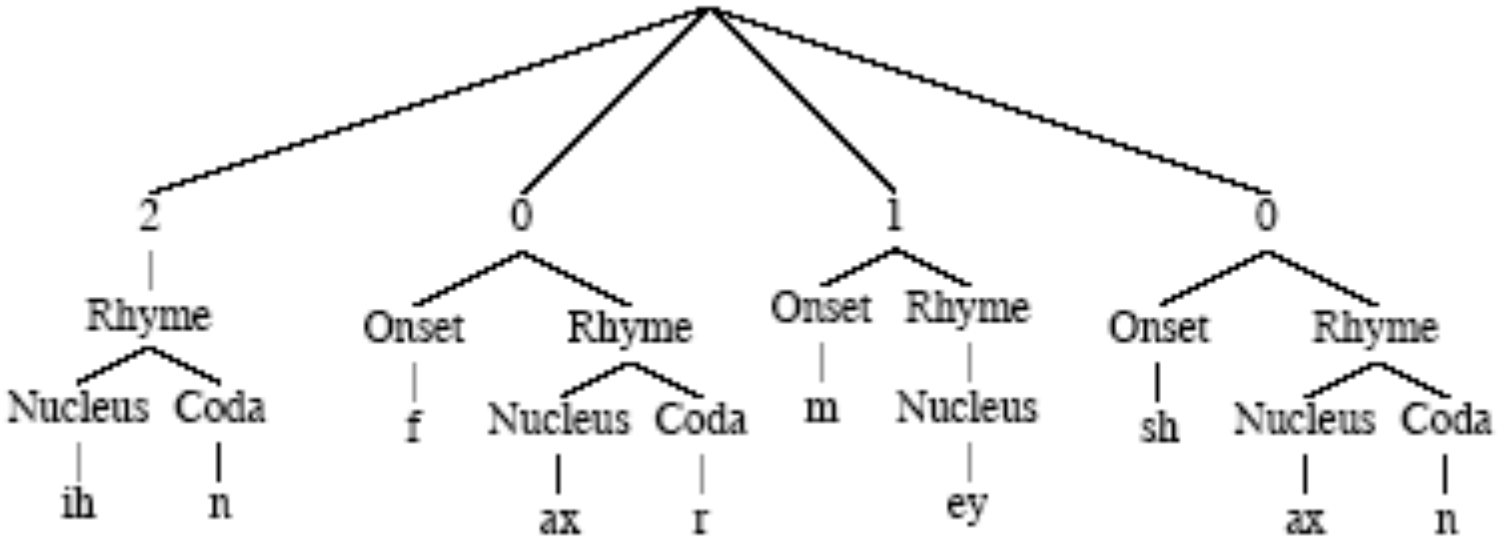
More phonetic structure

- 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