Algorithms for NLP

Speech Signals

Taylor Berg-Kirkpatrick – CMU

Slides: Dan Klein – UC Berkeley

Log-linear Parameterization

- Model form:

$$P(y|x;w) = \frac{\exp(w^\top f(x,y))}{\sum_{y'} \exp(w^\top f(x,y'))}$$

- Learn by following gradient of training LL:

$$\frac{\partial L(w)}{\partial w} = \sum_i f(x_i, y_i^*) - \sum_i \left(\mathbb{E}_{P(y|x_i;w)} \left[f(x_i, y) \right] \right)$$

Mixed Interpolation

- But can’t we just interpolate:
  - P(w|most recent words)
  - P(w|skip contexts)
  - P(w|caching)
  - …
- Yes, and people do (well, did)
  - But additive combination tends to flatten distributions, not zero out candidates

Neural LMs

Neural LMs

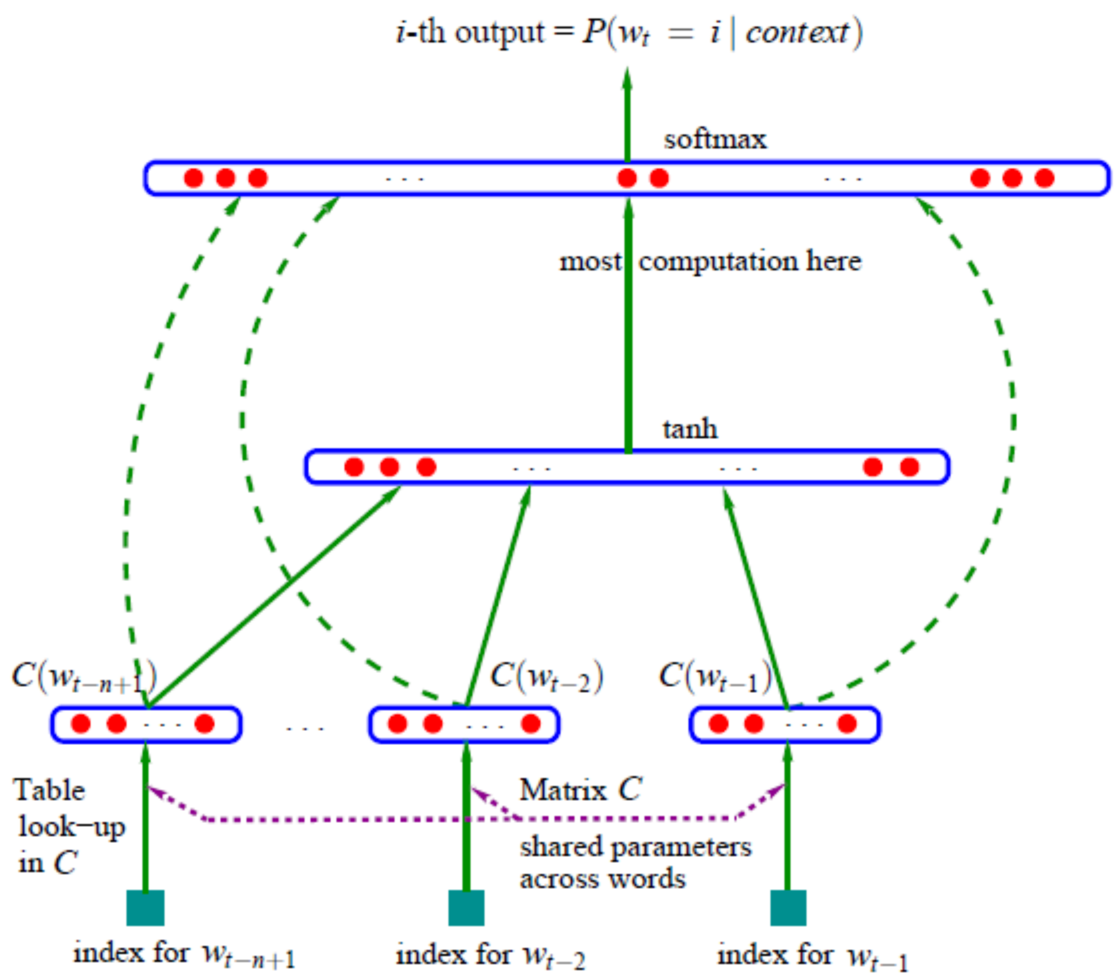


Image: (Bengio et al, 03)

Neural vs Maxent

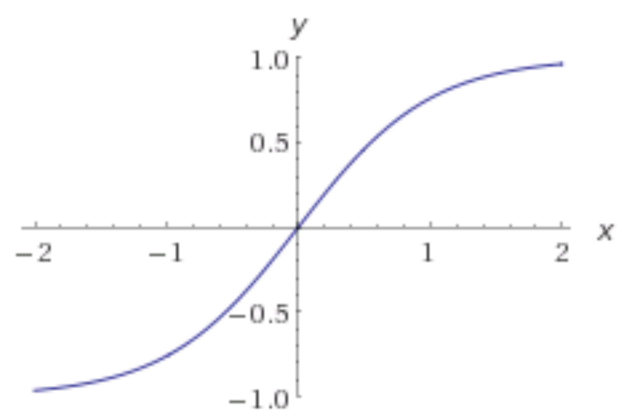
- Maxent LM

$$P(y|x;w) \propto \exp(w^\top f(x,y))$$

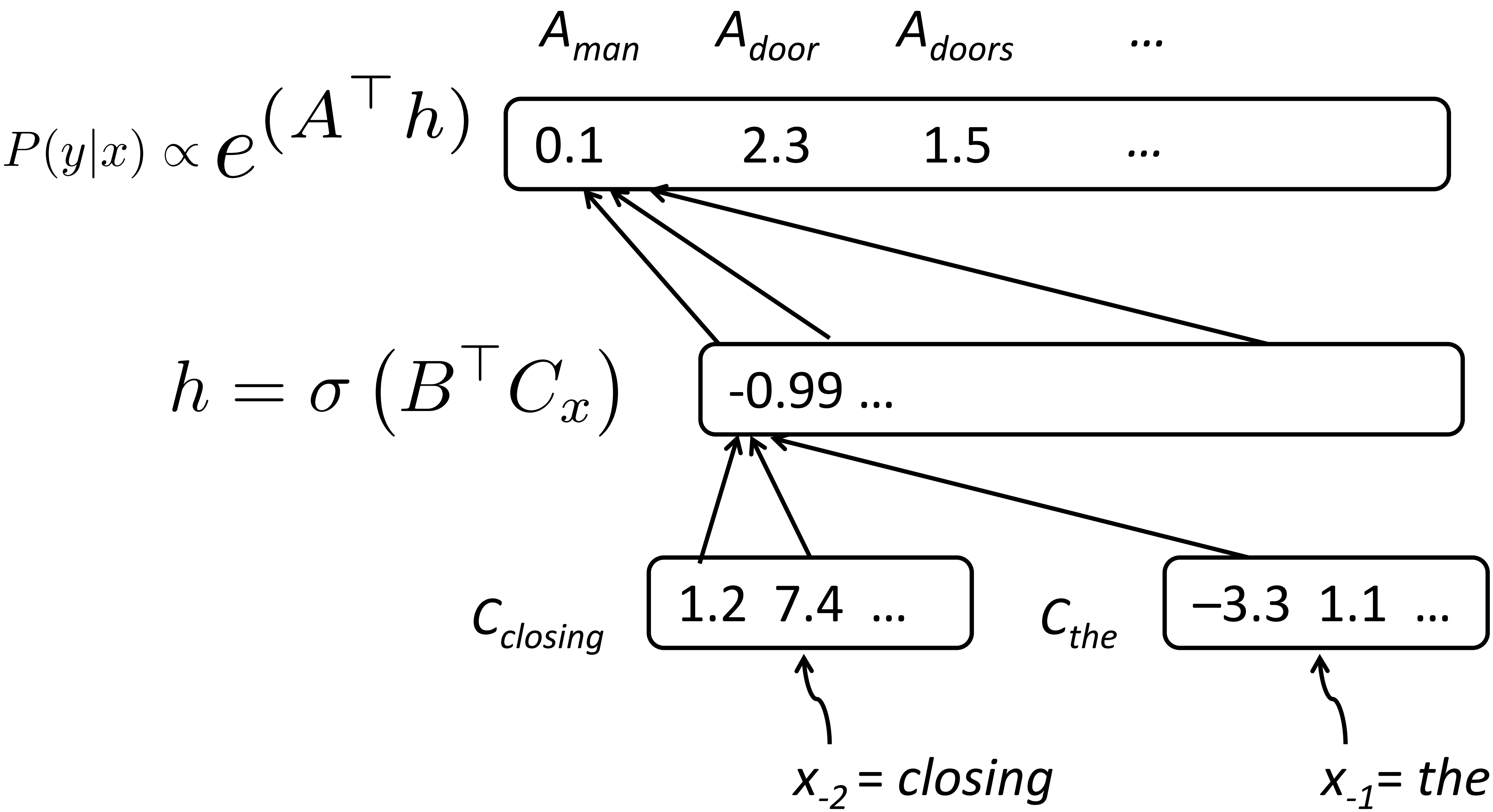
- Simple Neural LM

$$P(y|x;A,B,C) \propto \exp\left(A_y^\top \sigma(B^\top C_x)\right)$$

σ nonlinear, e.g. tanh



Neural LM Example



Neural LMs

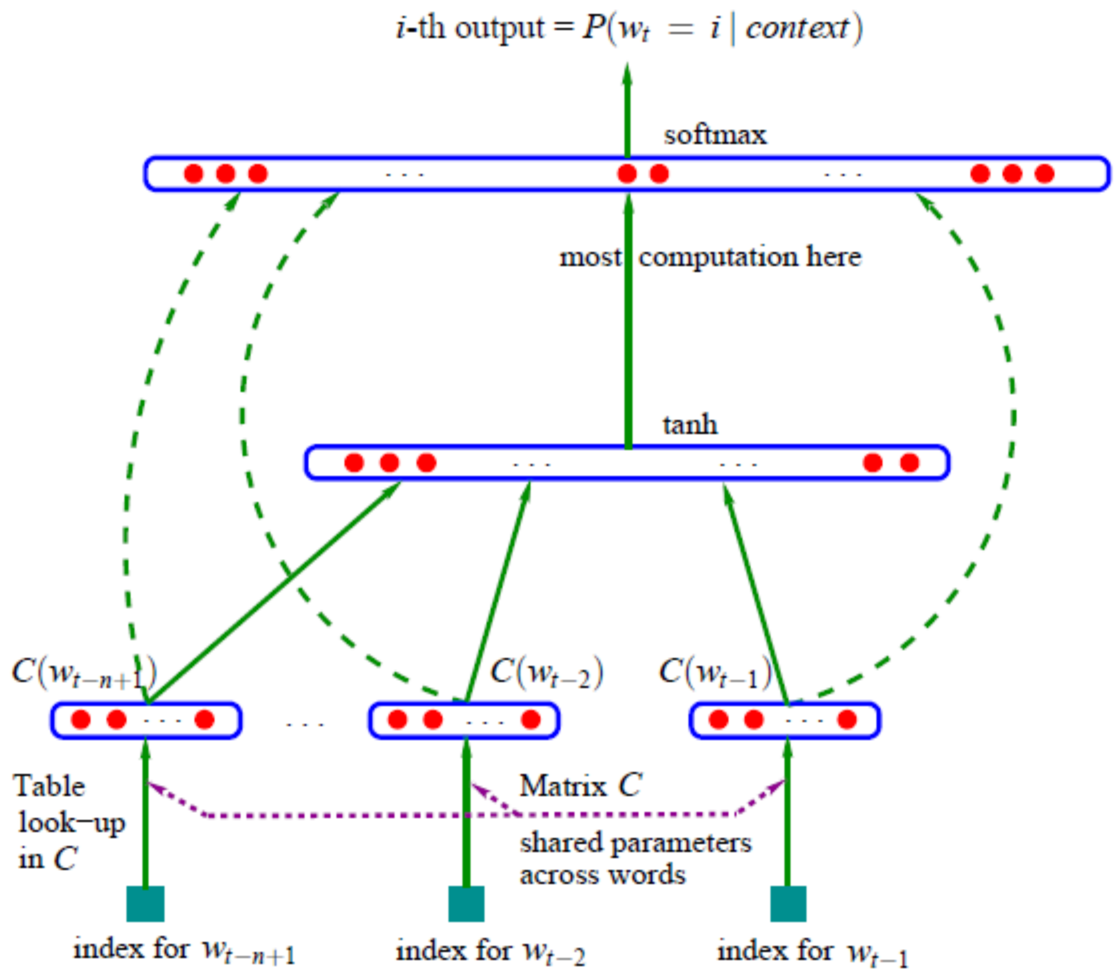
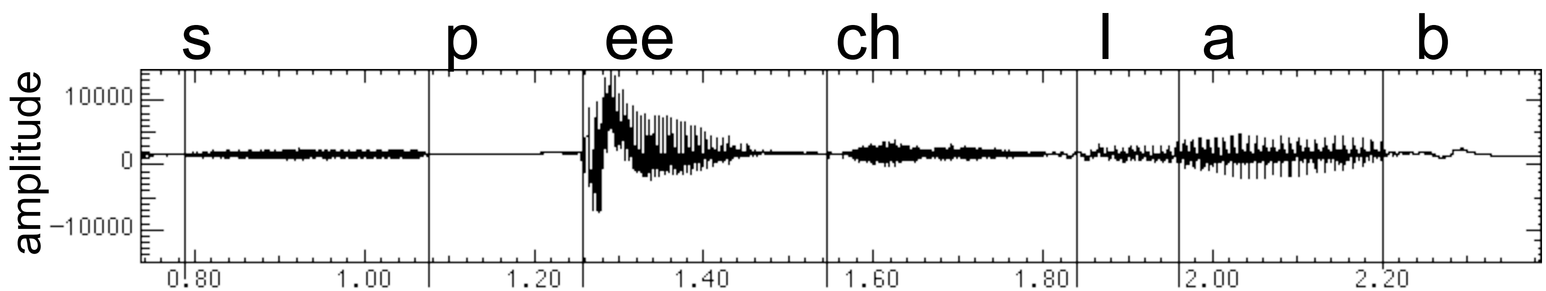


Image: (Bengio et al, 03)

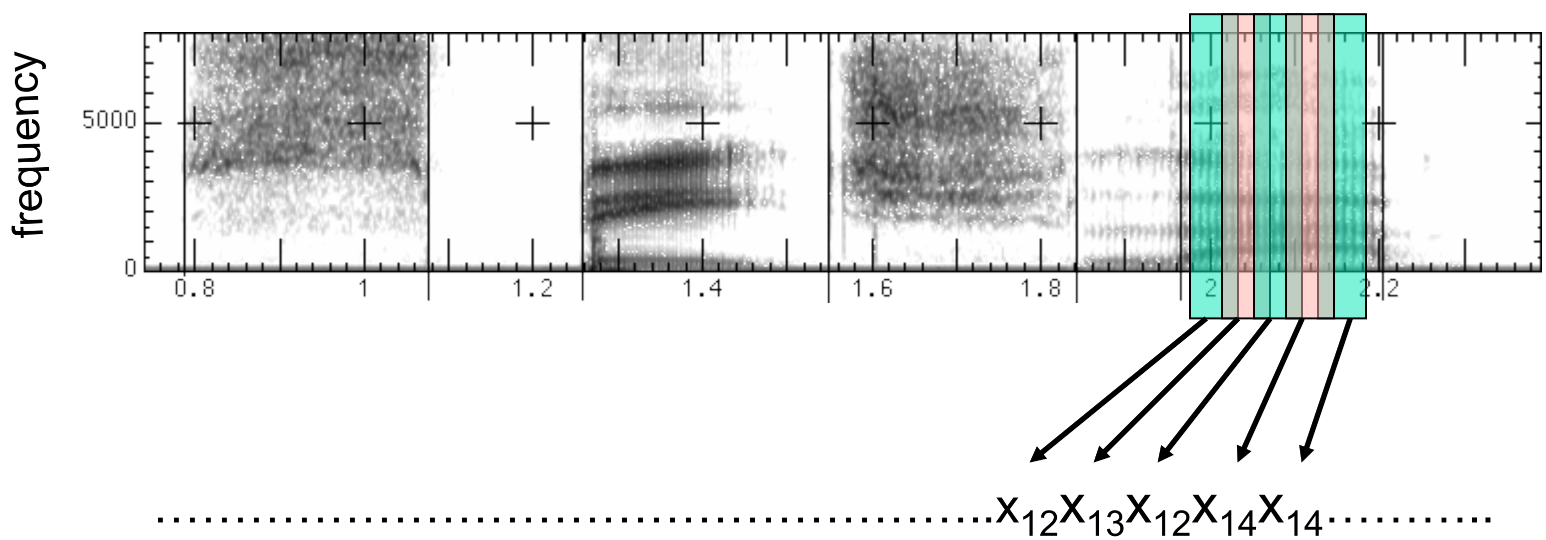
Speech Signals

Speech in a Slide

- Frequency gives pitch; amplitude gives volume

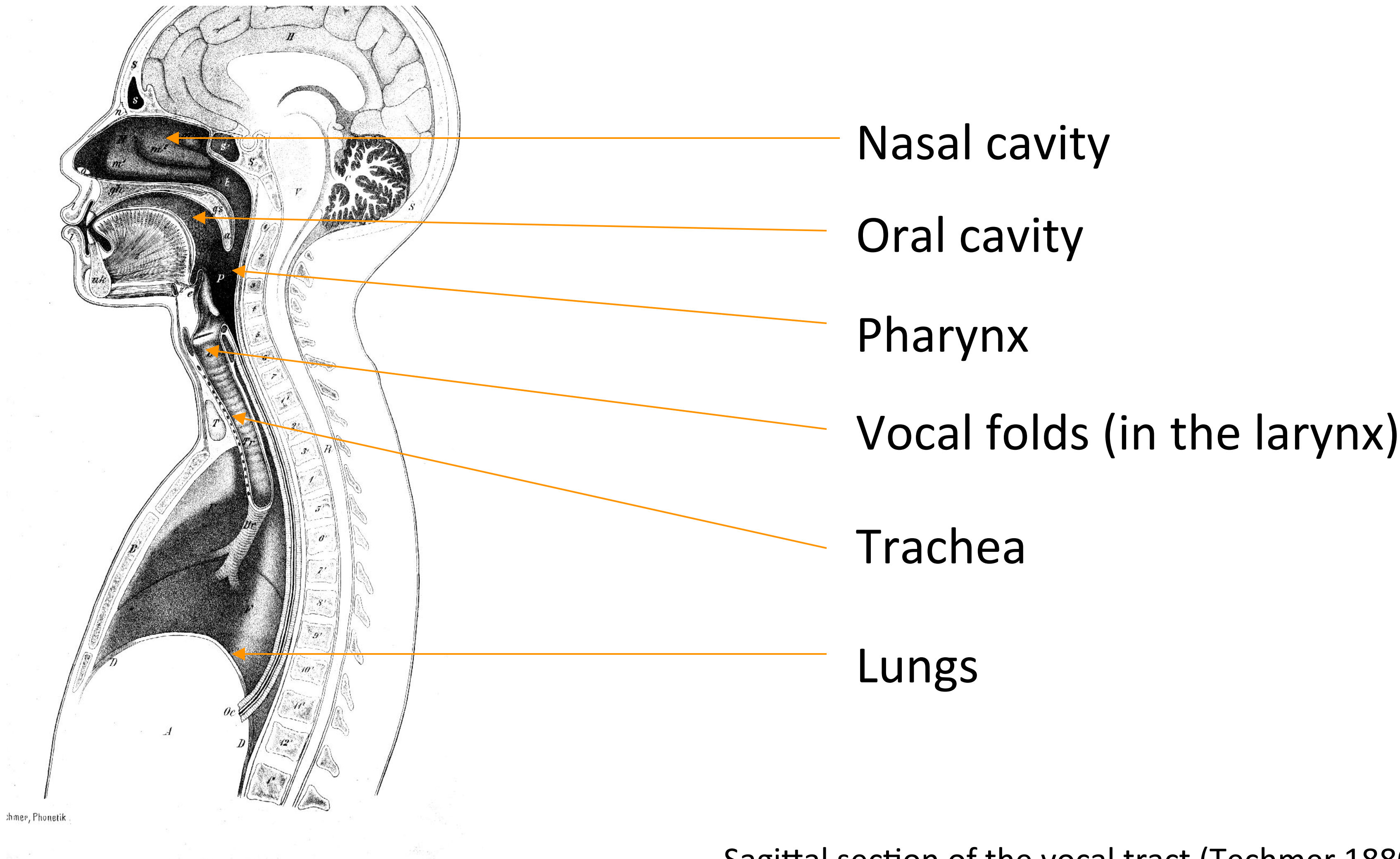


- Frequencies at each time slice processed into observation vectors



Articulation

Articulatory System



Sagittal section of the vocal tract (Techmer 1880)

Text from Ohala, Sept 2001, from Sharon Rose slide

Space of Phonemes

	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		я	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ		ꞎ	𝼆	𝼄				
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		𝼈						

- Standard international phonetic alphabet (IPA) chart of consonants

Place

Places of Articulation

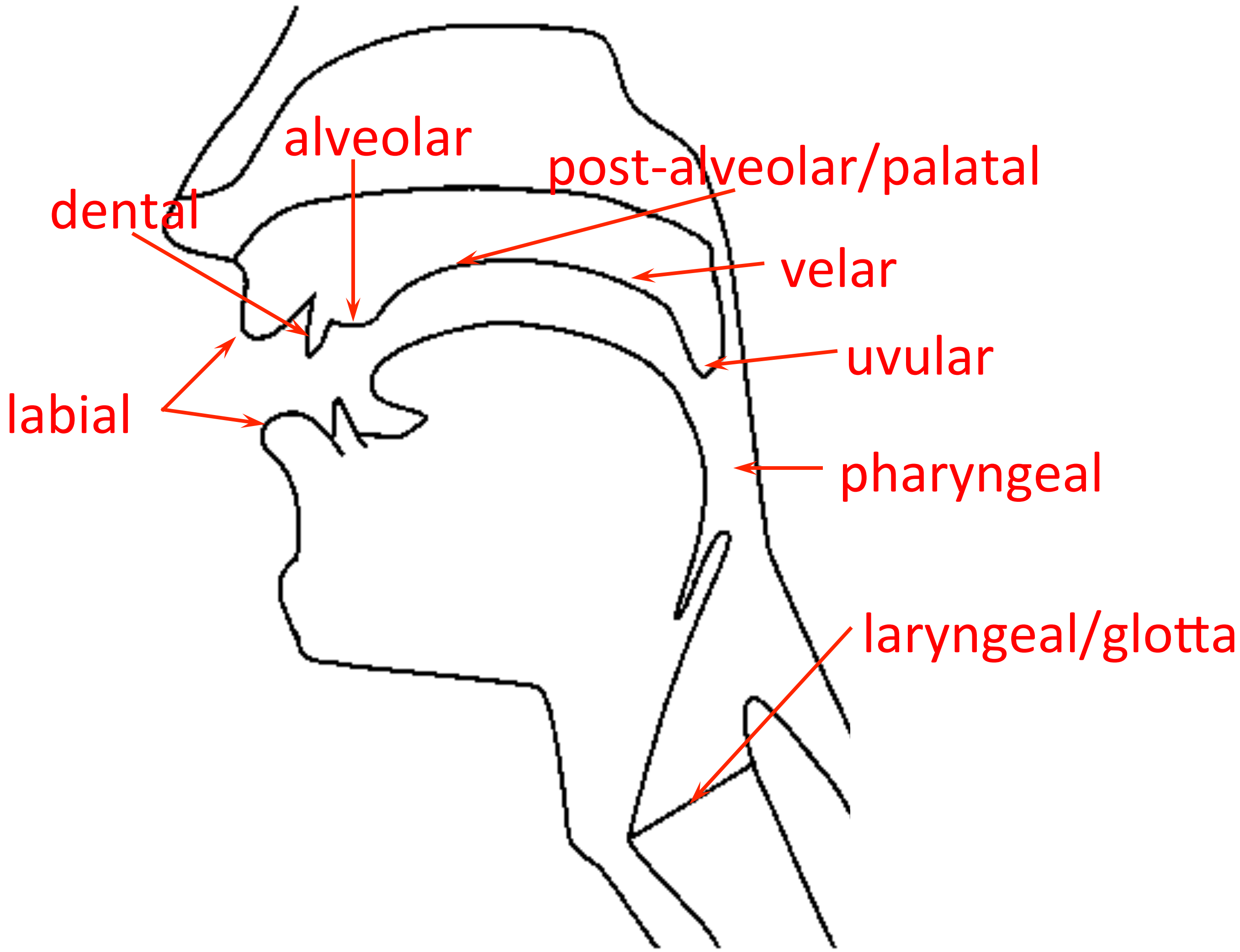
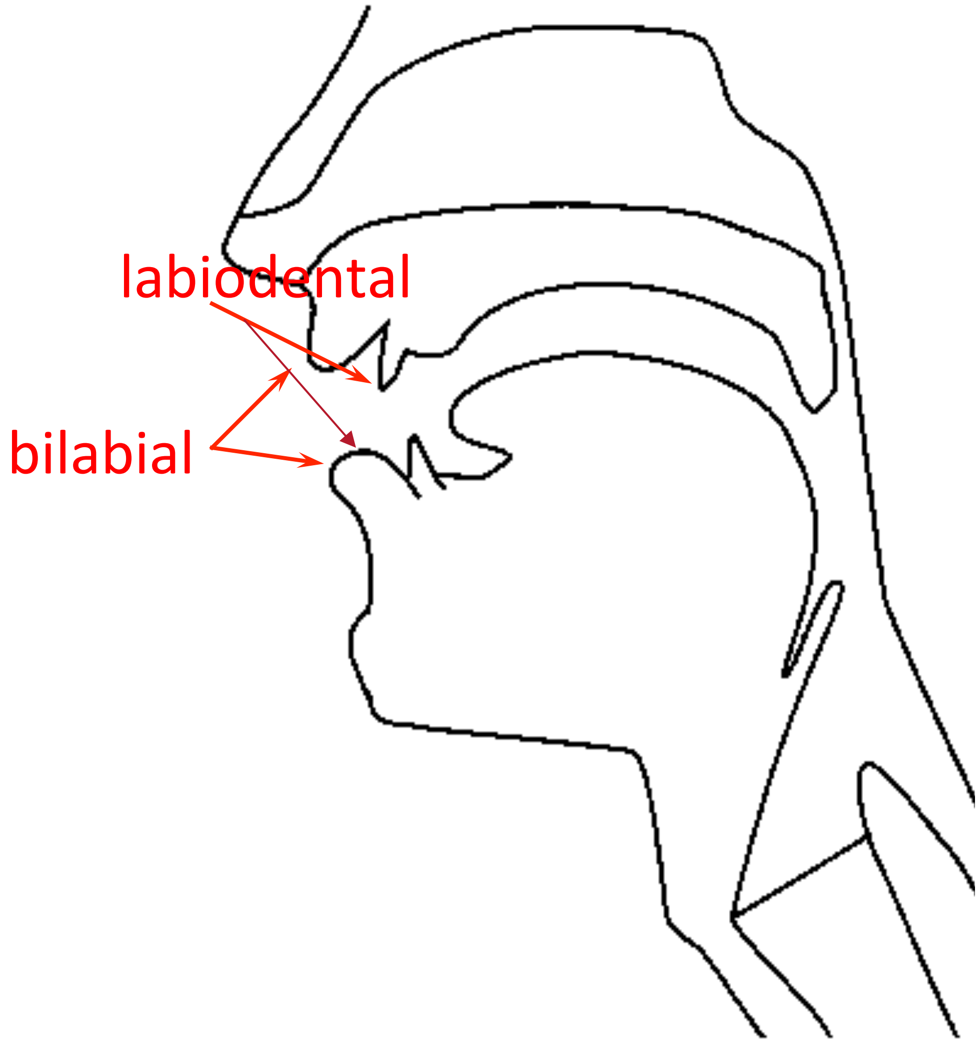


Figure thanks to Jennifer Venditti

Labial place

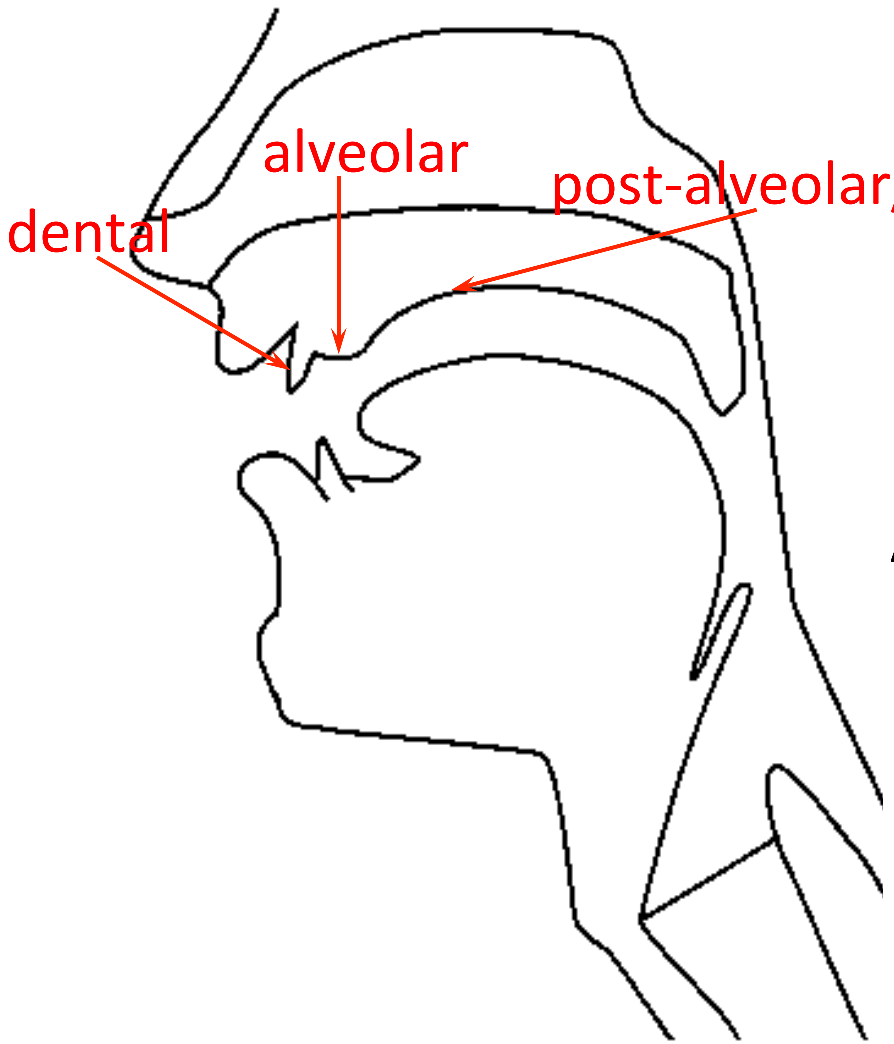


Bilabial:
p, b, m

Labiodental:
f, v

Figure thanks to Jennifer Venditti

Coronal place



Dental:

th/dh

Alveolar:

t/d/s/z/l/n

Post:

sh/zh/y

Figure thanks to Jennifer Venditti

Dorsal Place

Velar:

k/g/ng

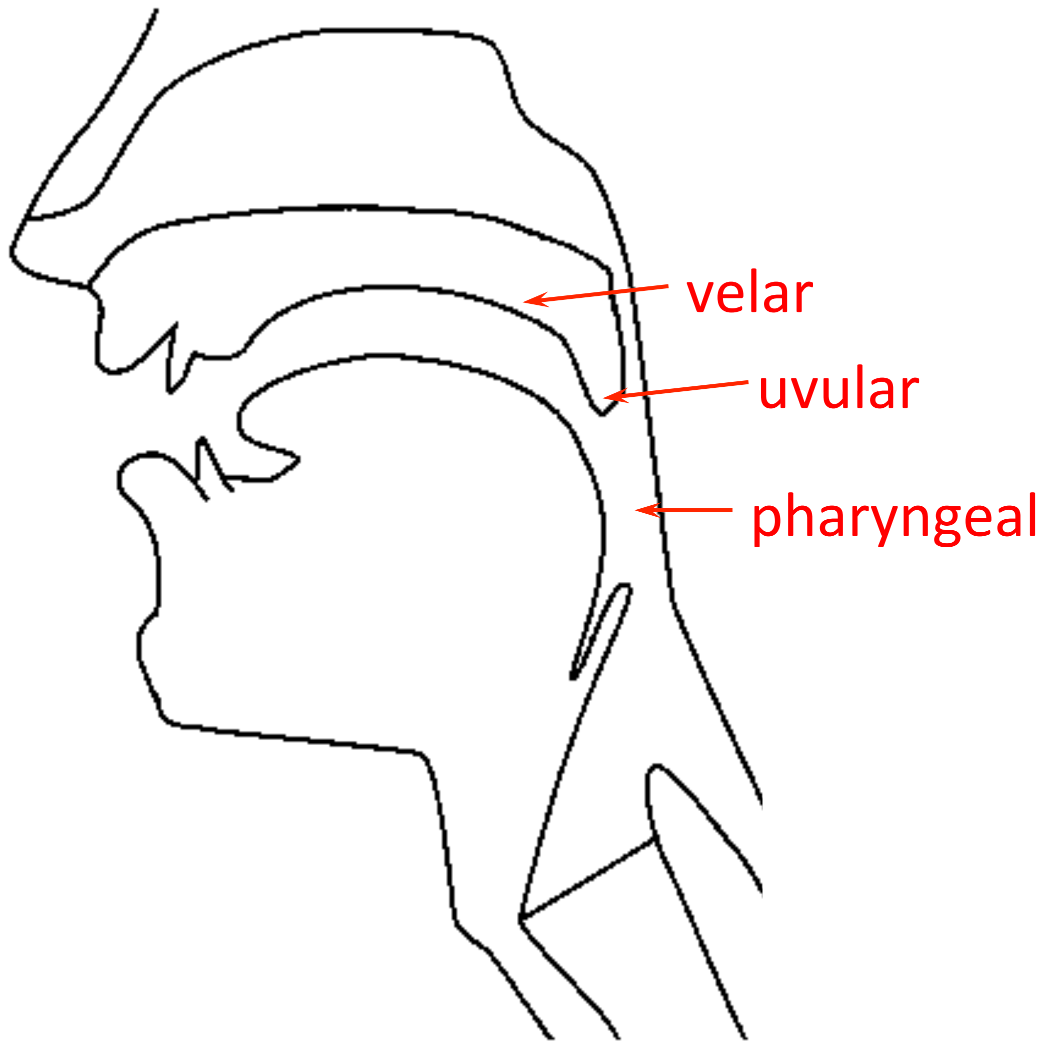


Figure thanks to Jennifer Venditti

Space of Phonemes

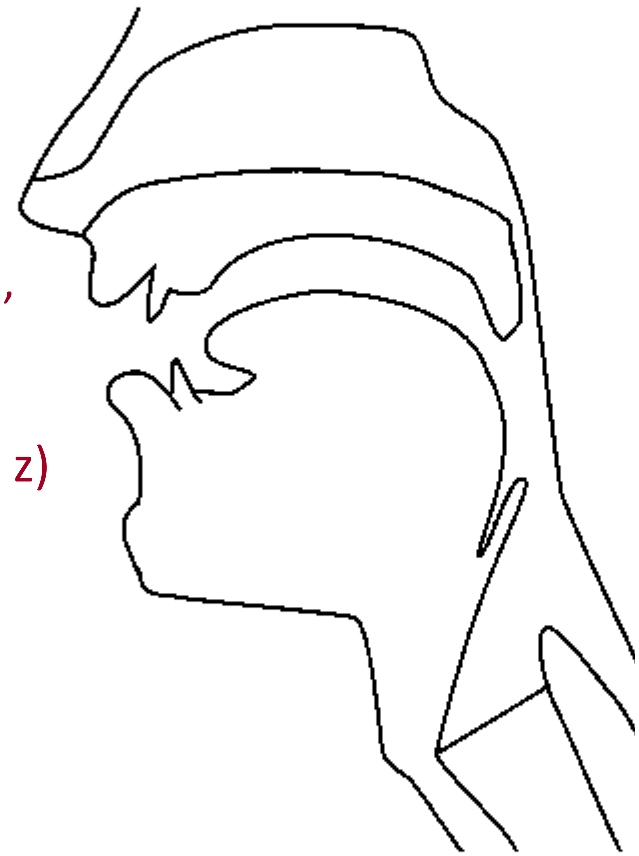
	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		ᴙ	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ		ꞎ	𝼆	𝼄				
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		𝼈						

- Standard international phonetic alphabet (IPA) chart of consonants

Manner

Manner of Articulation

- In addition to varying by place, sounds vary by manner
- Stop: complete closure of articulators, no air escapes via mouth
  - Oral stop: palate is raised (p, t, k, b, d, g)
  - Nasal stop: oral closure, but palate is lowered (m, n, ng)
- Fricatives: substantial closure, turbulent: (f, v, s, z)
- Approximants: slight closure, sonorant: (l, r, w)
- Vowels: no closure, sonorant: (i, e, a)



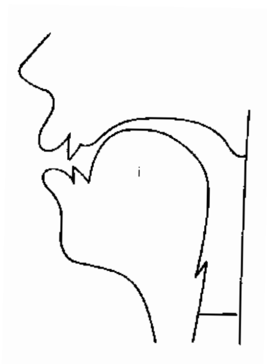
Space of Phonemes

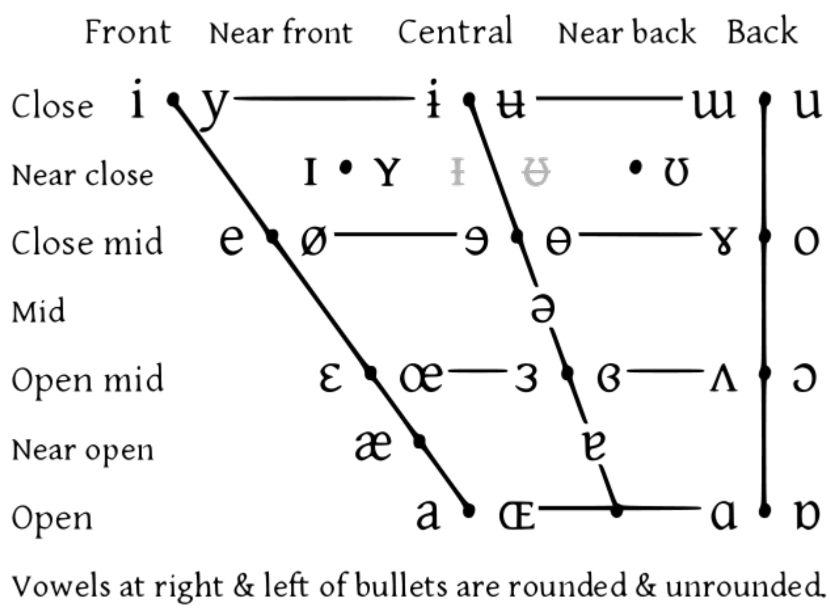
	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		я	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ		ꞎ	ʎ̝	ʟ̝				
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		ɺ̢						

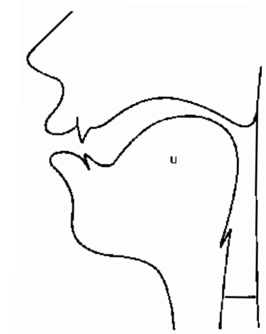
- Standard international phonetic alphabet (IPA) chart of consonants

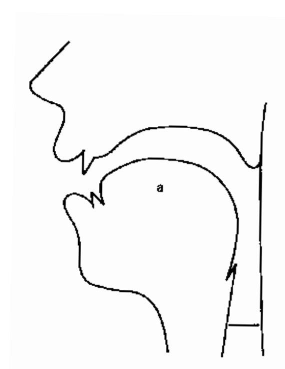
Vowels

Vowel Space



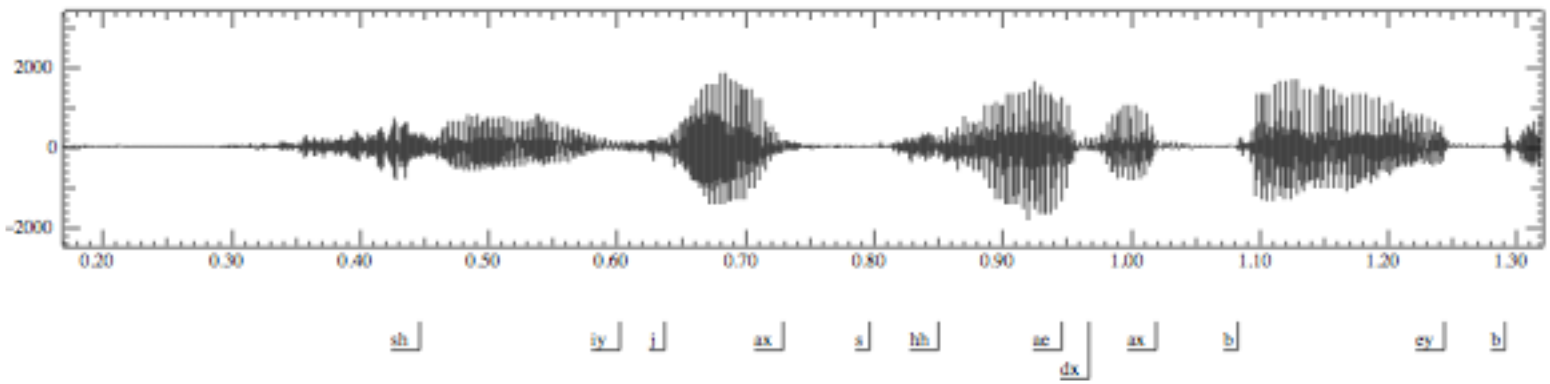

Front Near front Central Near back Back
Close i y ɨ ʉ ɯ u
Near close ɪ ʏ ᵻ ᵿ ʊ
Close mid e ø ɘ ɵ ɤ o
Mid ə
Open mid ɛ œ ɜ ɞ ʌ ɔ
Near open æ ɐ
Open a ɶ ɑ ɒ
Vowels at right & left of bullets are rounded & unrounded.





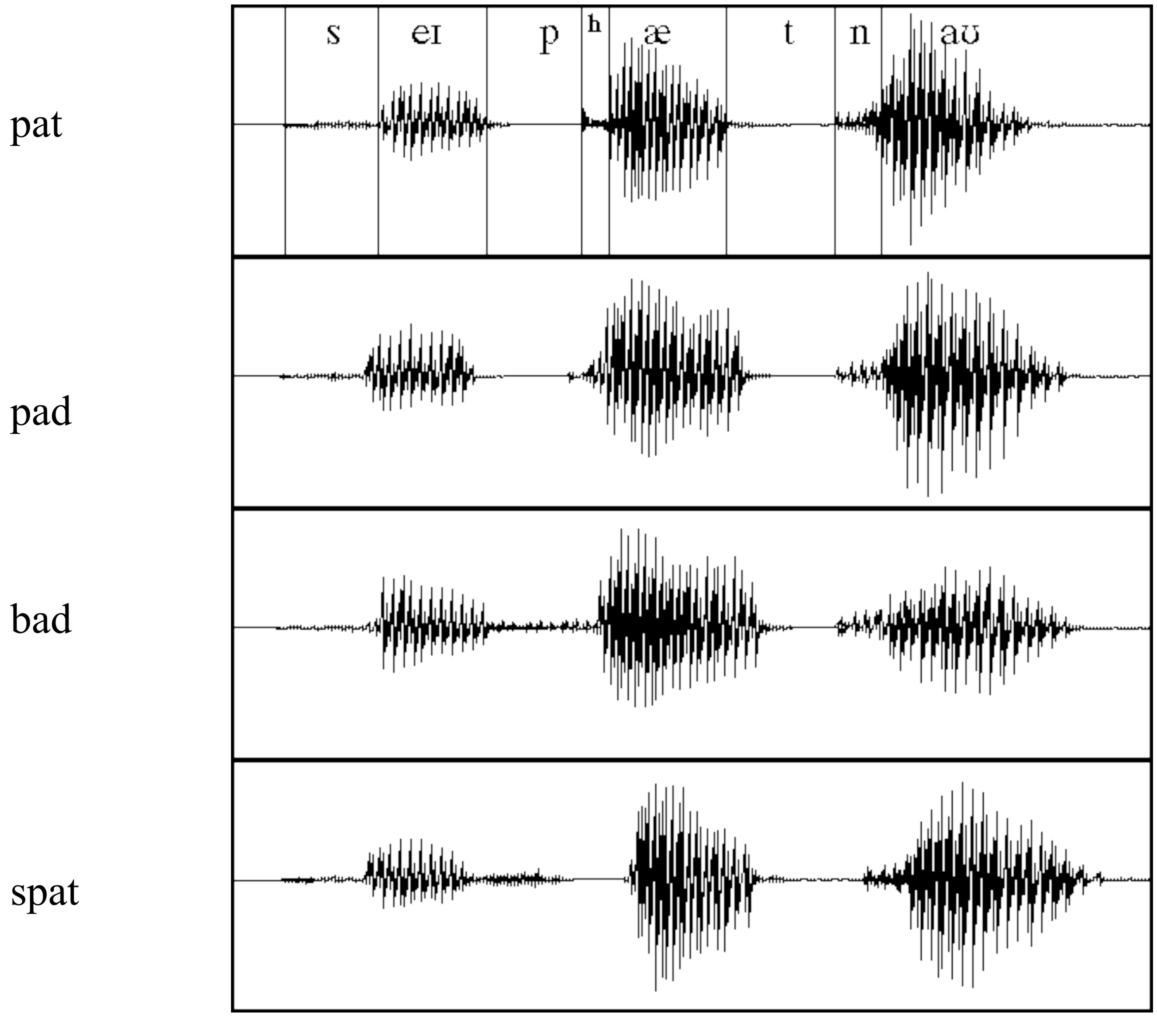
Acoustics

"She just had a baby"



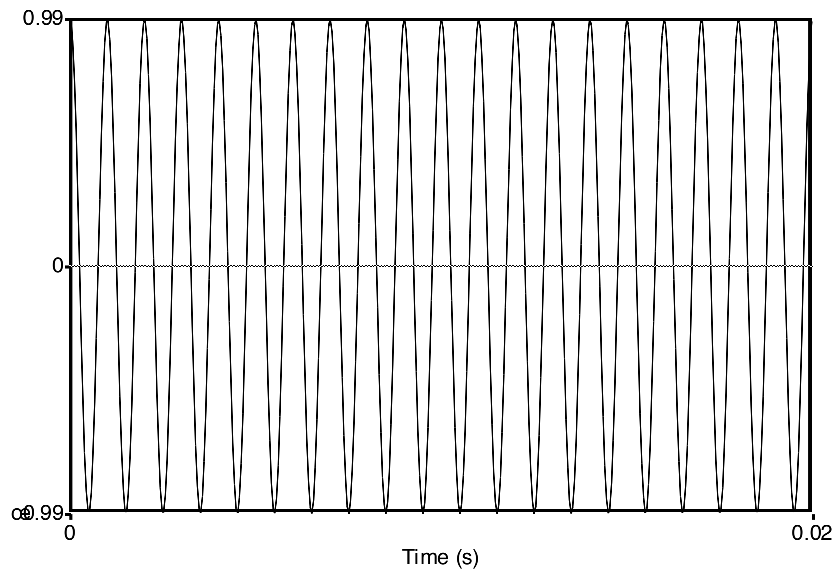
- What can we learn from a wavefile?
  - No gaps between words (!)
  - Vowels are voiced, long, loud
  - Length in time = length in space in waveform picture
  - Voicing: regular peaks in amplitude
  - When stops closed: no peaks, silence
  - Peaks = voicing: .46 to .58 (vowel [iy], from second .65 to .74 (vowel [ax]) and so on
  - Silence of stop closure (1.06 to 1.08 for first [b], or 1.26 to 1.28 for second [b])
  - Fricatives like [sh]: intense irregular pattern; see .33 to .46

Time-Domain Information



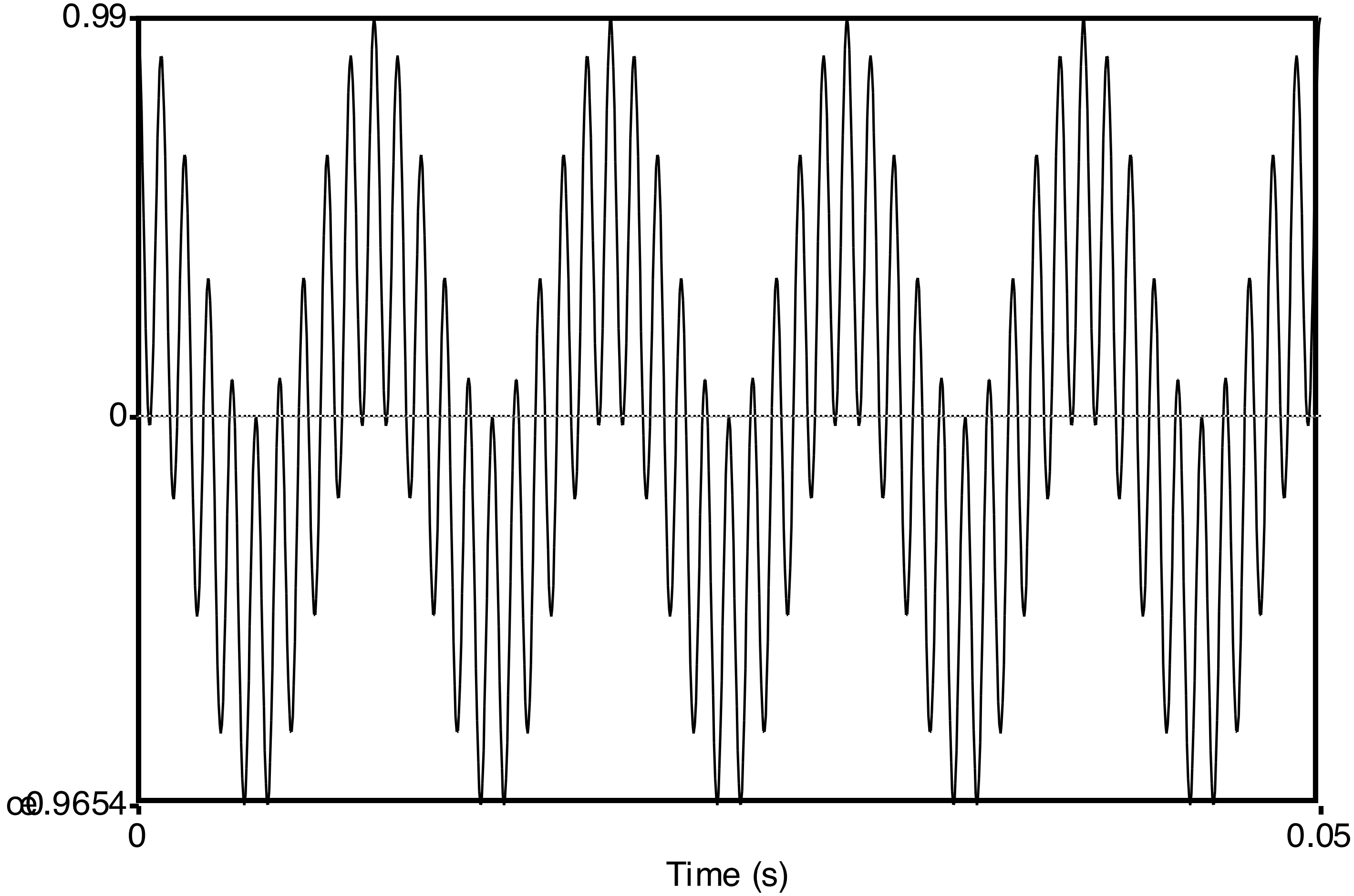
Example from Ladefoged

Simple Periodic Waves of Sound



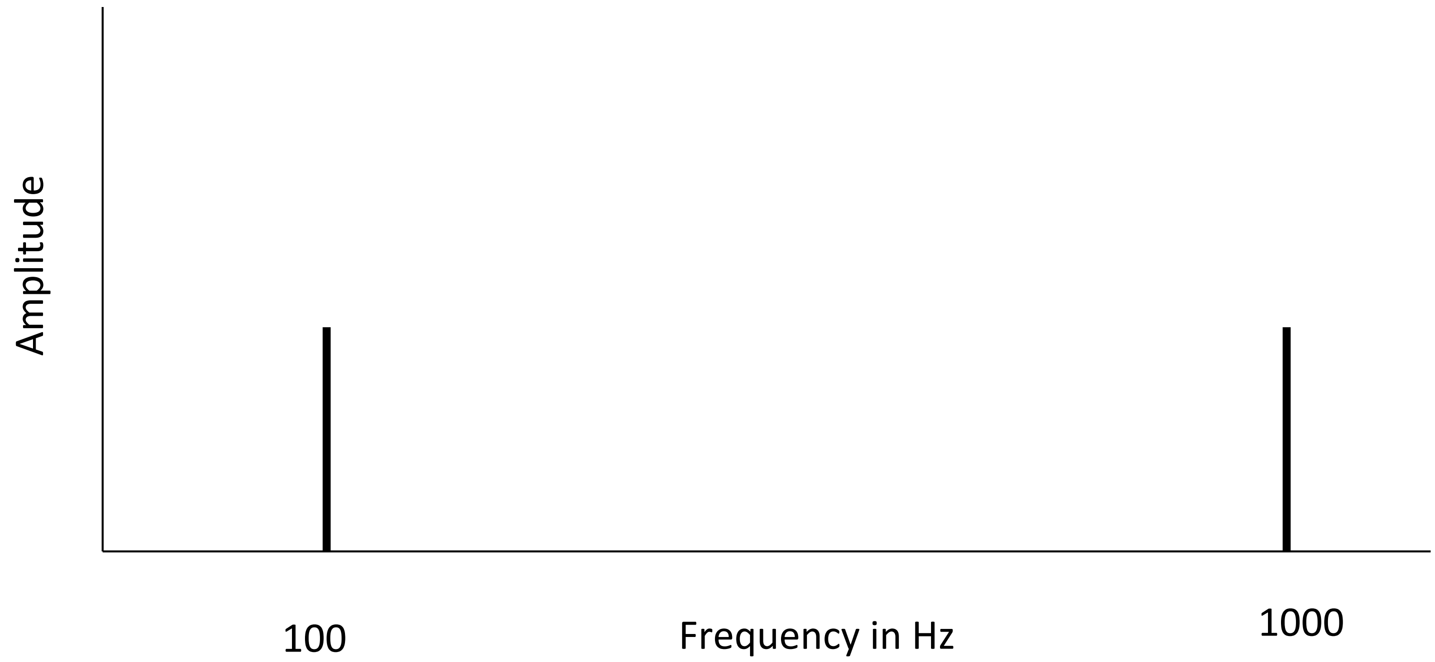
- Y axis: Amplitude = amount of air pressure at that point in time
  - Zero is normal air pressure, negative is rarefaction
- X axis: Time.
- Frequency = number of cycles per second.
- 20 cycles in .02 seconds = 1000 cycles/second = 1000 Hz

Complex Waves: 100Hz+1000Hz

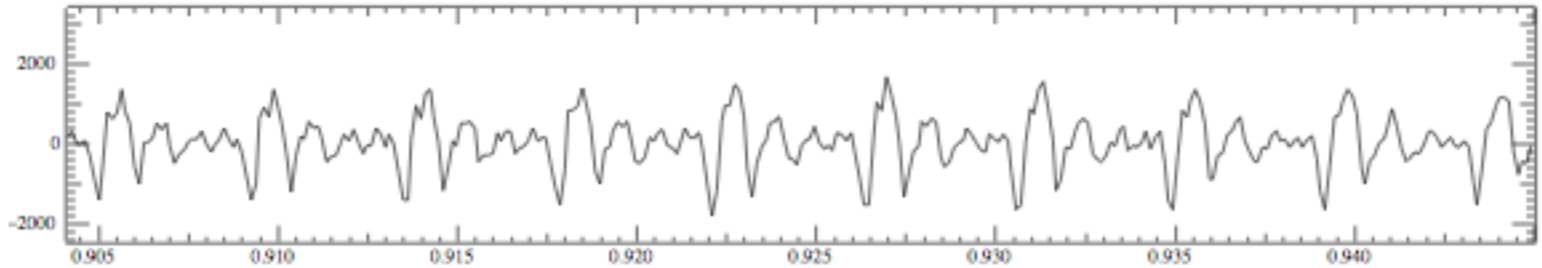


Spectrum

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

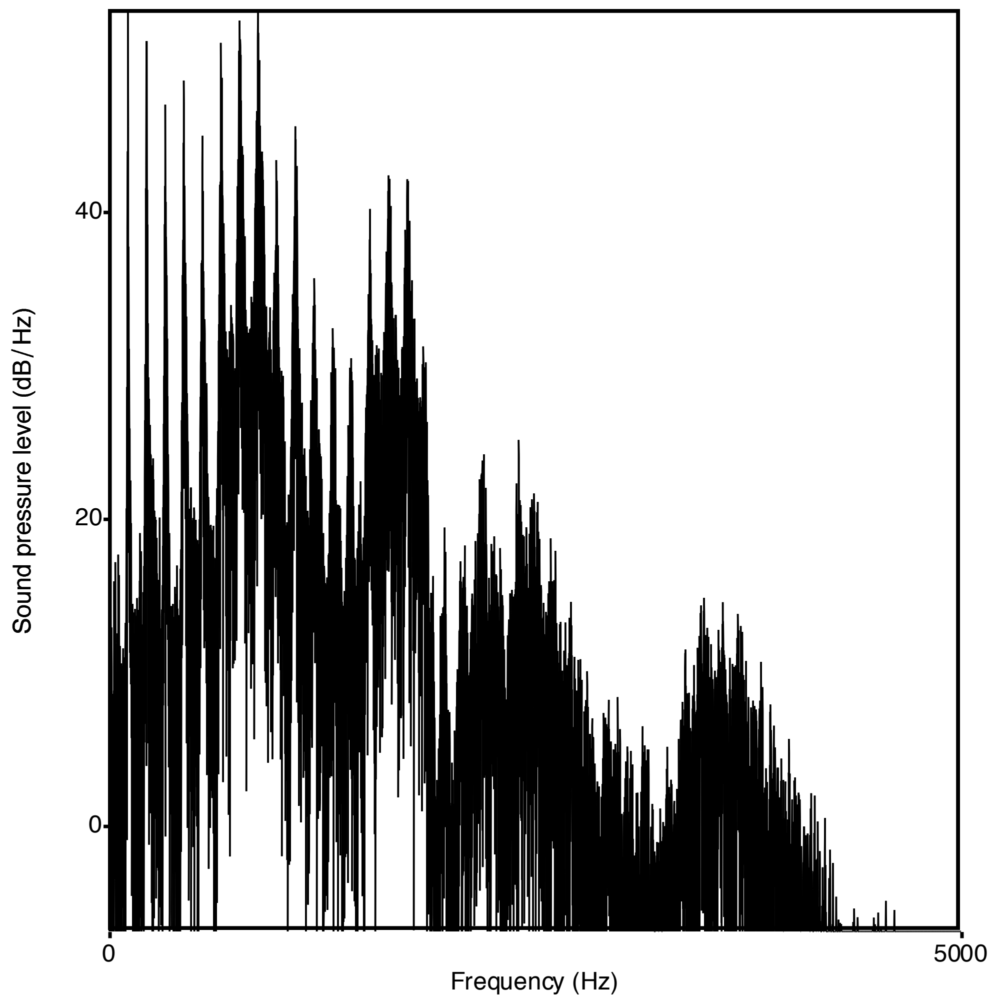


Part of [ae] waveform from "had"



- 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

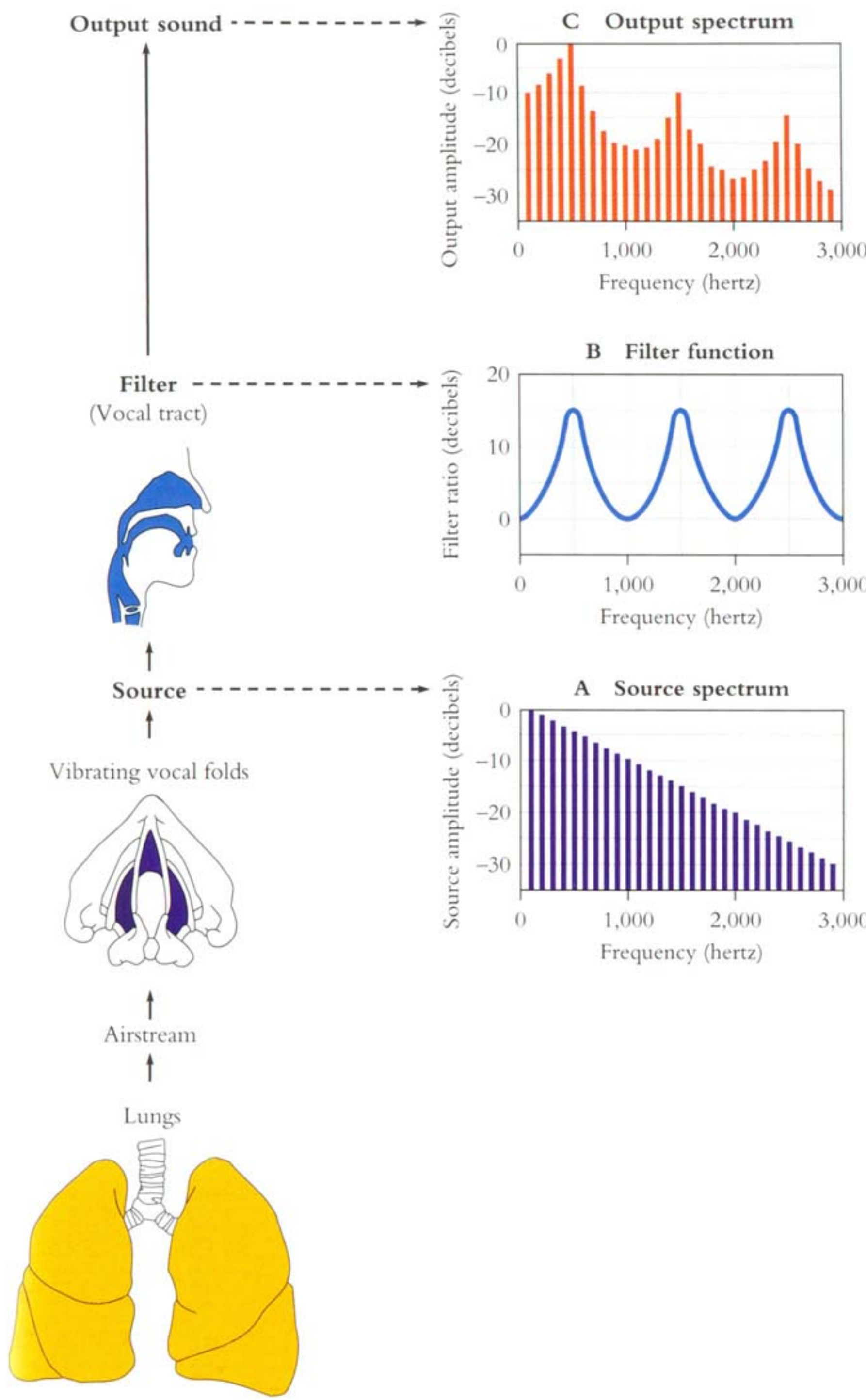
Spectrum of an Actual Soundwave



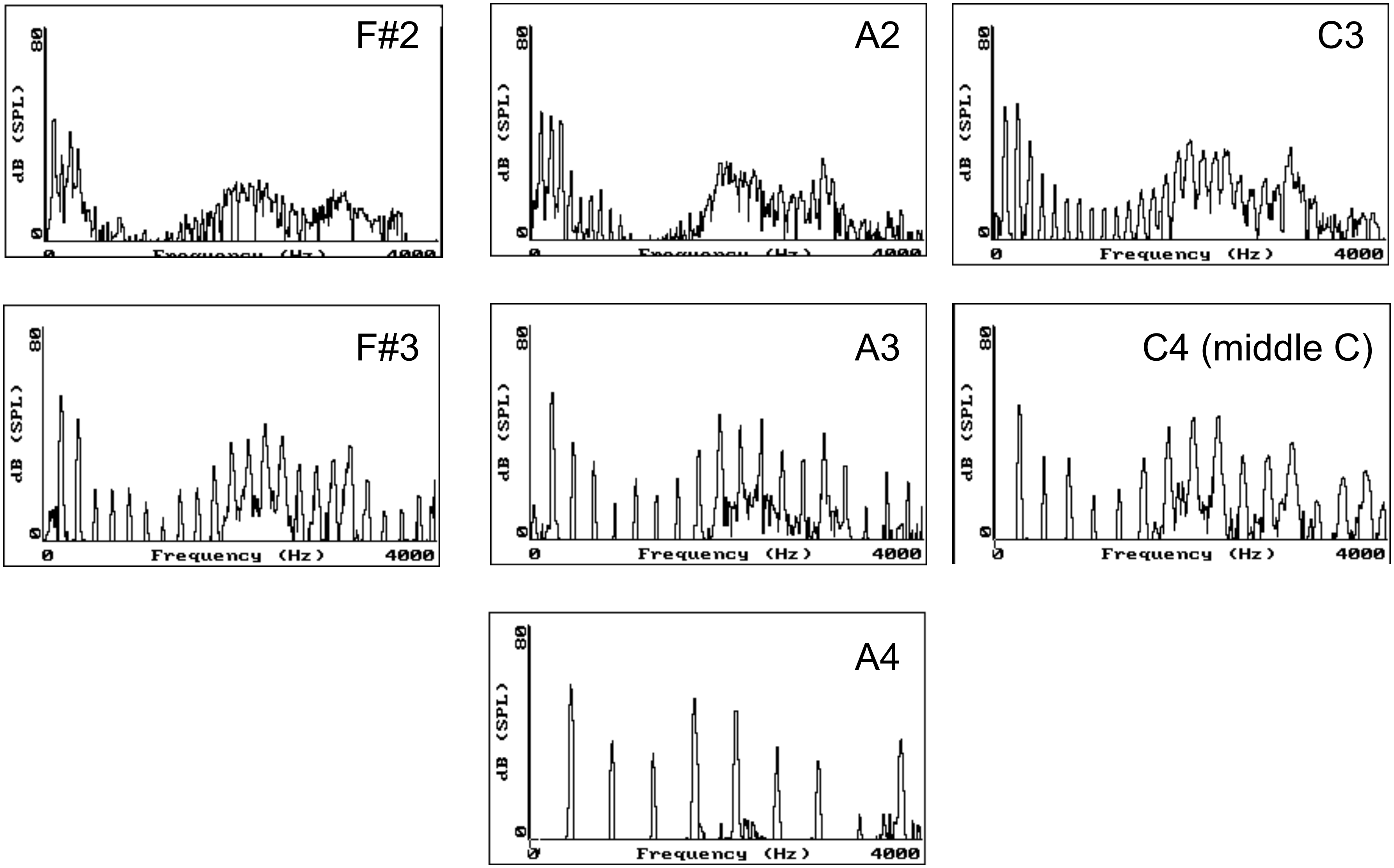
Source / Channel

Why these Peaks?

- Articulation process:
  - The vocal cord vibrations create harmonics
  - The mouth is an amplifier
  - Depending on shape of mouth, some harmonics are amplified more than others



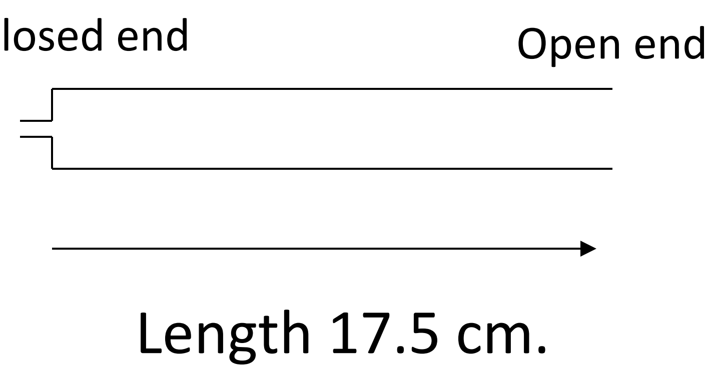
Vowel [i] at increasing pitches



Figures from Ratree Wayland

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.
- Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end.

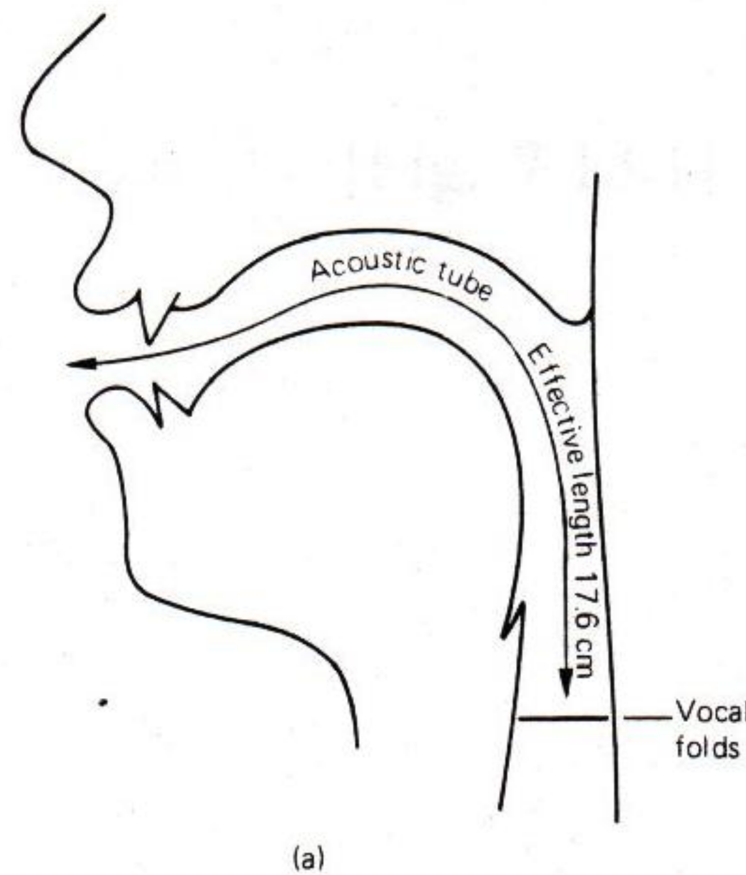
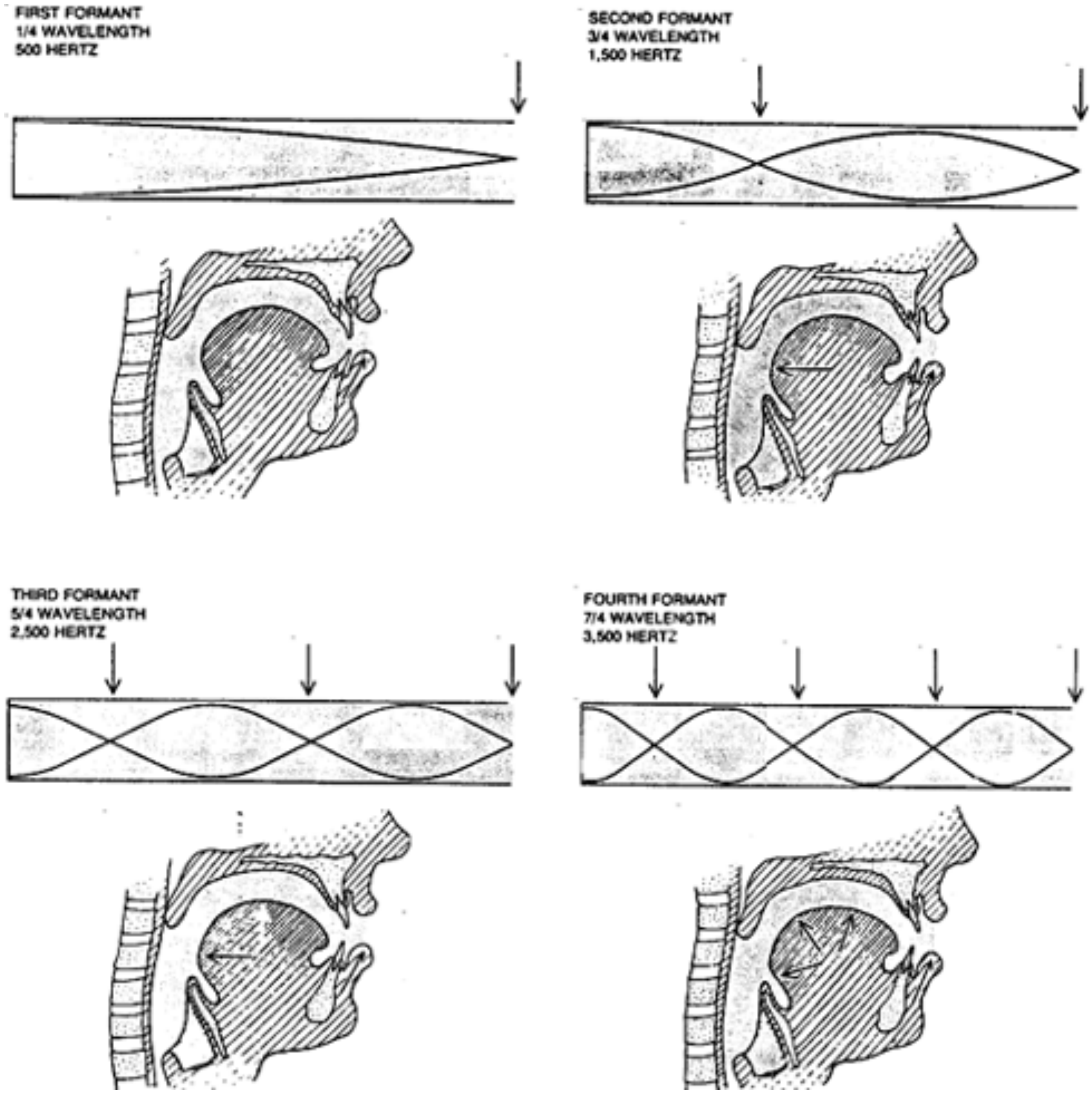


Figure from W. Barry



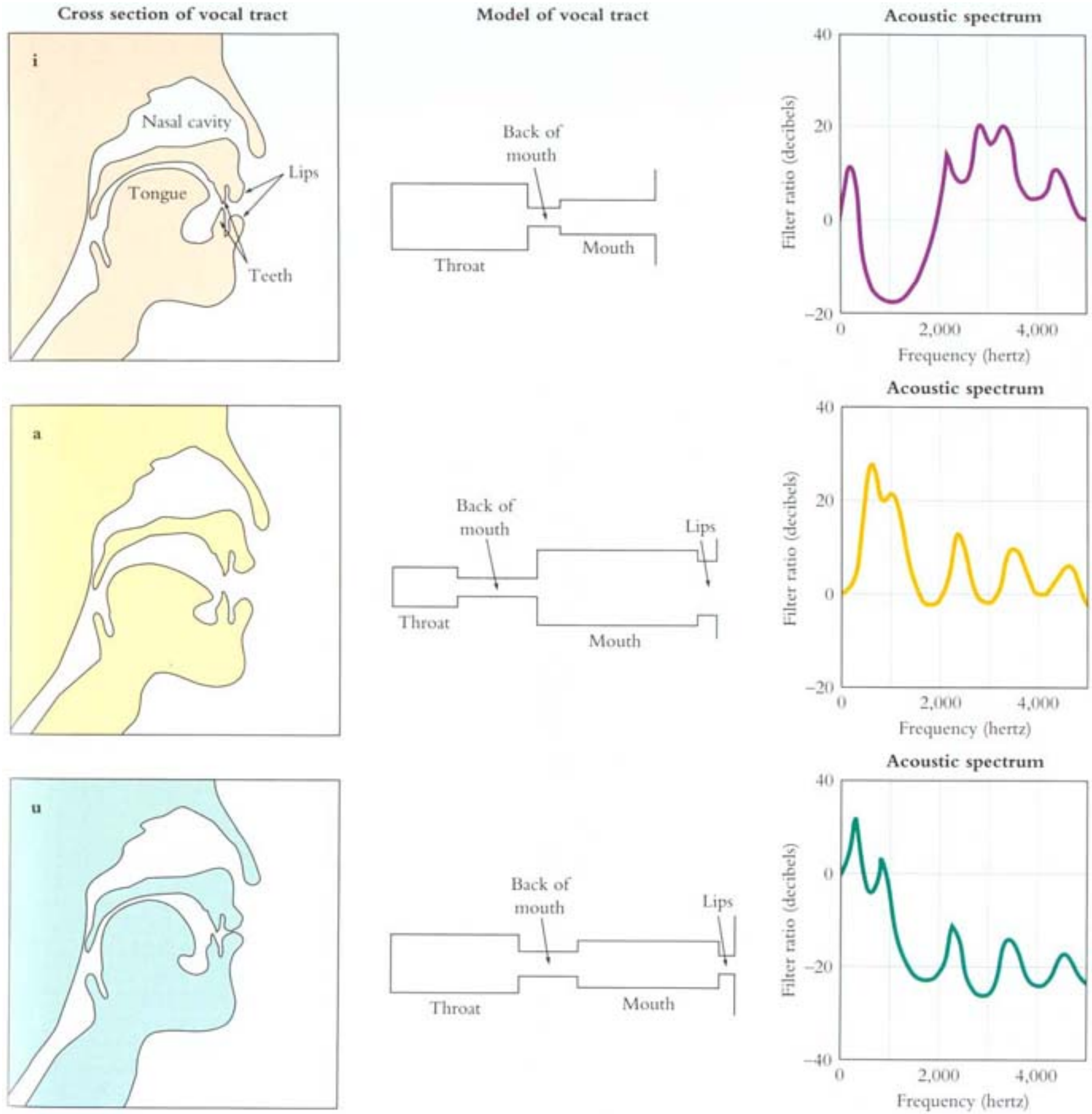
From Sundberg

Computing the 3 Formants of Schwa

- Let the length of the tube be L
  - $F_1 = c/\lambda_1 = c/(4L) = 35{,}000/4*17.5 =$ 500Hz
  - $F_2 = c/\lambda_2 = c/(4/3L) = 3c/4L = 3*35{,}000/4*17.5 =$ 1500Hz
  - $F_3 = c/\lambda_3 = c/(4/5L) = 5c/4L = 5*35{,}000/4*17.5 =$ 2500Hz

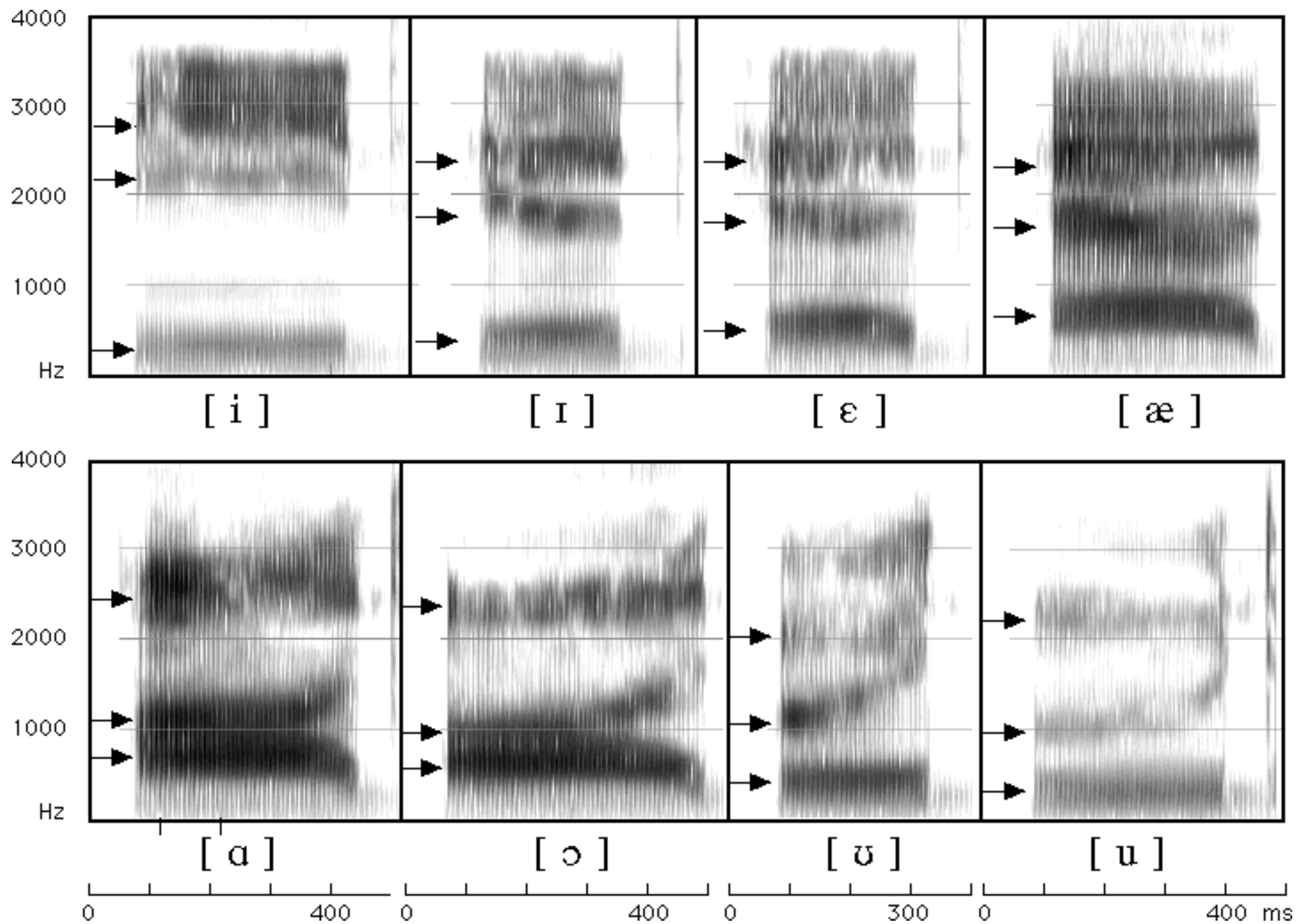
- So we expect a neutral vowel to have 3 resonances at 500, 1500, and 2500 Hz

- These vowel resonances are called formants

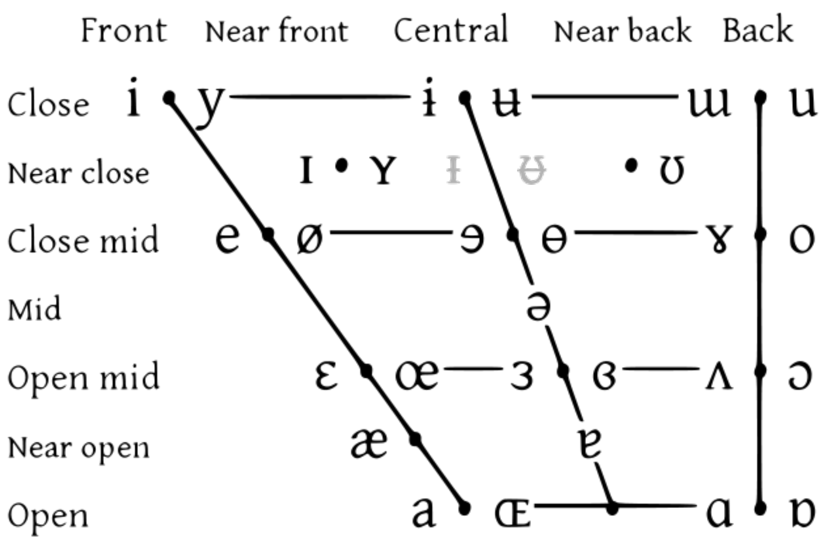


From Mark Liberman

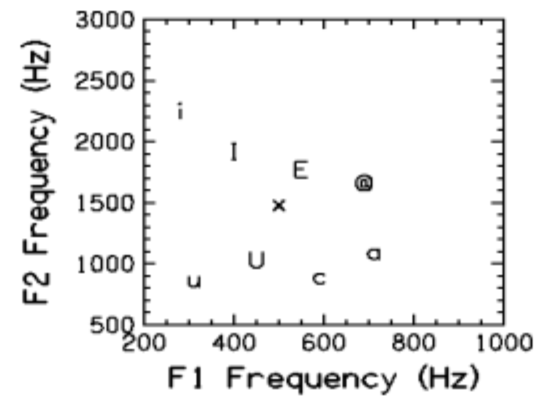
Seeing Formants: the Spectrogram

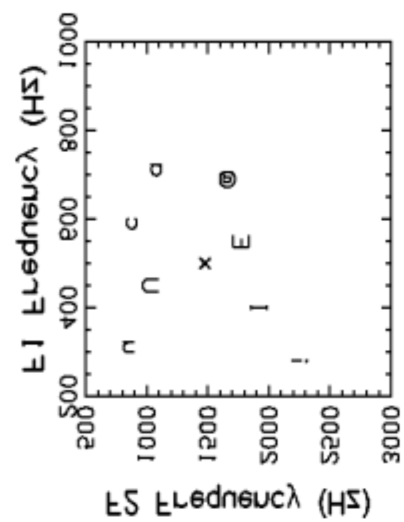


Vowel Space



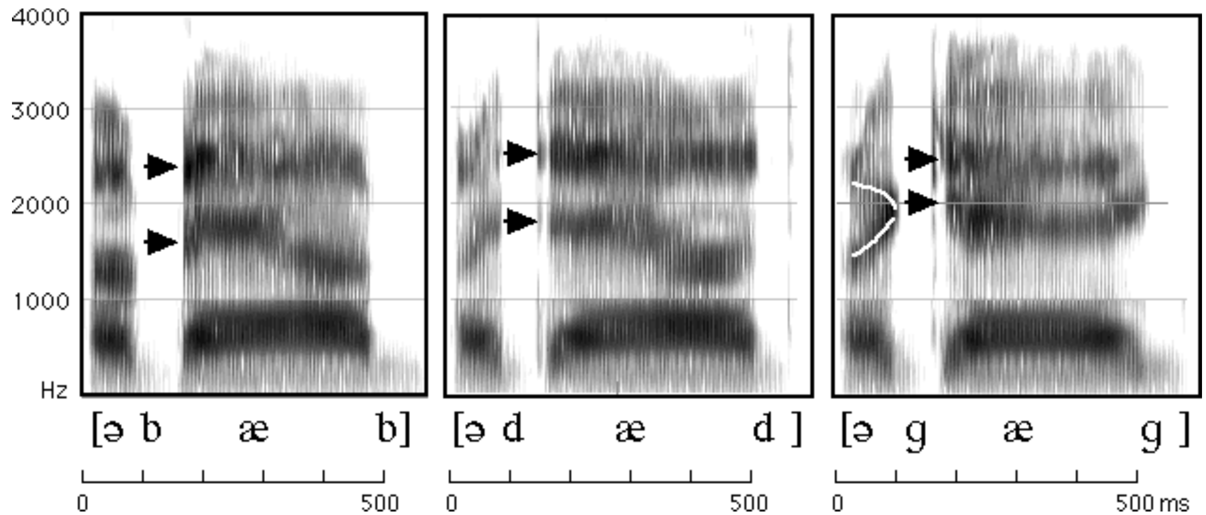
Vowels at right & left of bullets are rounded & unrounded.





Spectrograms

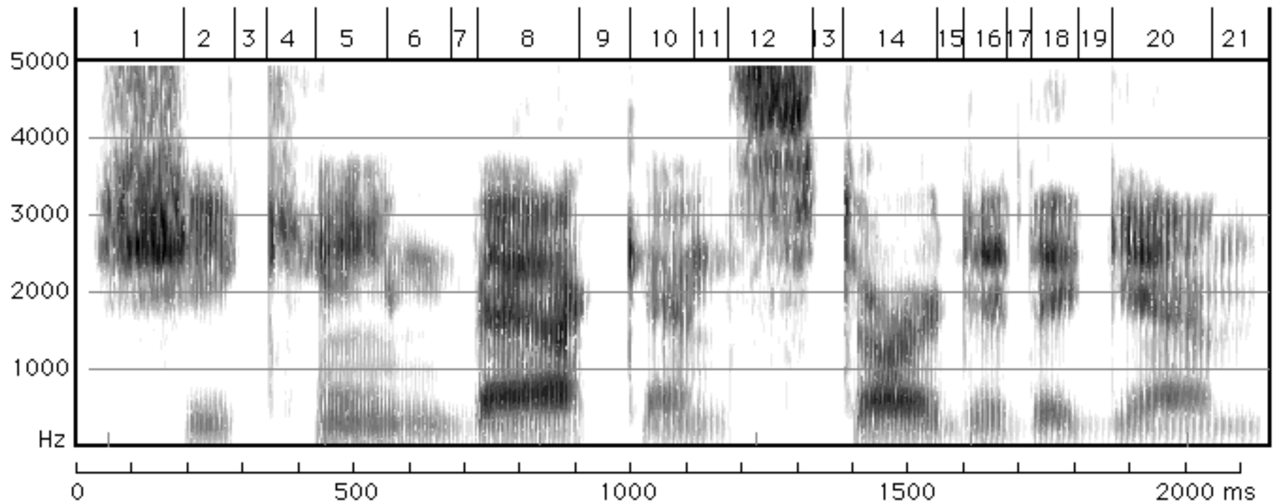
How to Read Spectrograms



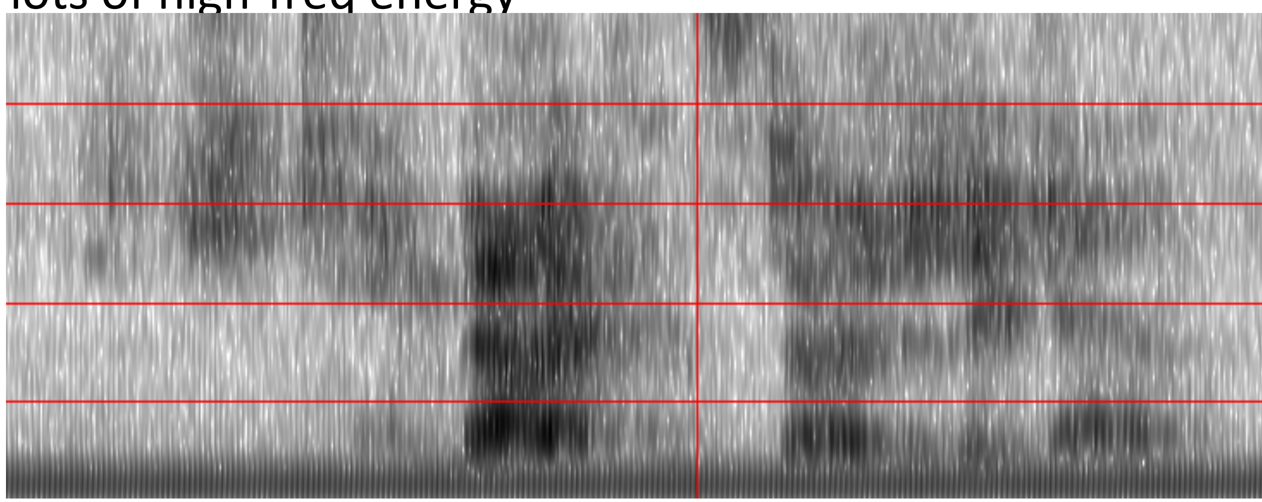
- [bab]: closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab"
- [dad]: first formant increases, but F2 and F3 slight fall
- [gag]: F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials

From Ladefoged "A Course in Phonetics"

“She came back and started again”



1. lots of high-freq energy



3.
4.
5.
6.
7.
8.
9.

From Ladefoged “A Course in Phonetics”