Algorithms for NLP

Speech Signals

Taylor Berg-Kirkpatrick – CMU

Slides: Dan Klein – UC Berkeley

Maximum Entropy Models

Improving on N-Grams?

- N-grams don't combine multiple sources of evidence well

P(construction | After the demolition was completed, the)

- Here:
  - "the" gives syntactic constraint
  - "demolition" gives semantic constraint
  - Unlikely the interaction between these two has been densely observed in this specific n-gram

- We'd like a model that can be more statistically efficient

Some Definitions

INPUTS	$\mathbf{x}_i$	*close the ____*
CANDIDATE SET	$\mathcal{Y}(\mathbf{x})$	*{door, table, …}*
CANDIDATES	$\mathbf{y}$	*table*
TRUE OUTPUTS	$\mathbf{y}_i^*$	*door*
FEATURE VECTORS	$\mathbf{f}(\mathbf{x}, \mathbf{y})$	[0 0 1 0 0 0 1 0 0 0 0 0]

x_{-1}="the" ∧ y="door"

x_{-1}="the" ∧ y="table"

"close" in x ∧ y="door"

y occurs in x

More Features, Less Interaction

x = closing the _____, y = doors

- N-Grams x_{-1}=*"the"* ∧ y="doors"
- Skips x_{-2}=*"closing"* ∧ y="doors"
- Lemmas x_{-2}=*"close"* ∧ y="door"
- Caching *y occurs in x*

Data: Feature Impact

Features	Train Perplexity	Test Perplexity
3 gram indicators	241	350
1-3 grams	126	172
1-3 grams + skips	101	164

Exponential Form

- Weights $\mathbf{w}$ Features $\mathbf{f}(\mathbf{x}, \mathbf{y})$
- Linear score $\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})$
- Unnormalized probability

$$\mathsf{P}(\mathbf{y}|\mathbf{x}, \mathbf{w}) \propto \mathsf{exp}(\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}))$$

- Probability

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

Likelihood Objective

- Model form:

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

- Log-likelihood of training data

$$L(w) = \log \prod_i P(y_i^*|x_i,w) = \sum_i \log \left(\frac{\exp(w^\top f(x_i,y_i^*))}{\sum_{y'} \exp(w^\top f(x_i,y'))} \right)$$

$$= \sum_i \left(w^\top f(x_i,y_i^*) - \log \sum_{y'} \exp(w^\top f(x_i,y')) \right)$$

Training

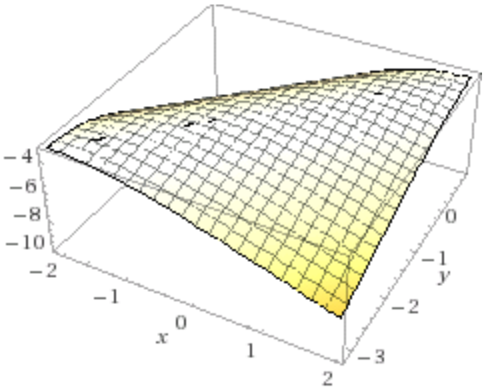
History of Training

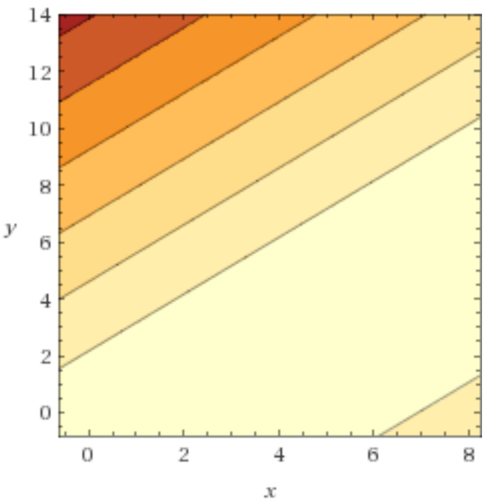
- 1990's: Specialized methods (e.g. iterative scaling)
- 2000's: General-purpose methods (e.g. conjugate gradient)
- 2010's: Online methods (e.g. stochastic gradient)

What Does LL Look Like?

- Example
  - Data: xxxy
  - Two outcomes, x and y
  - One indicator for each
  - Likelihood

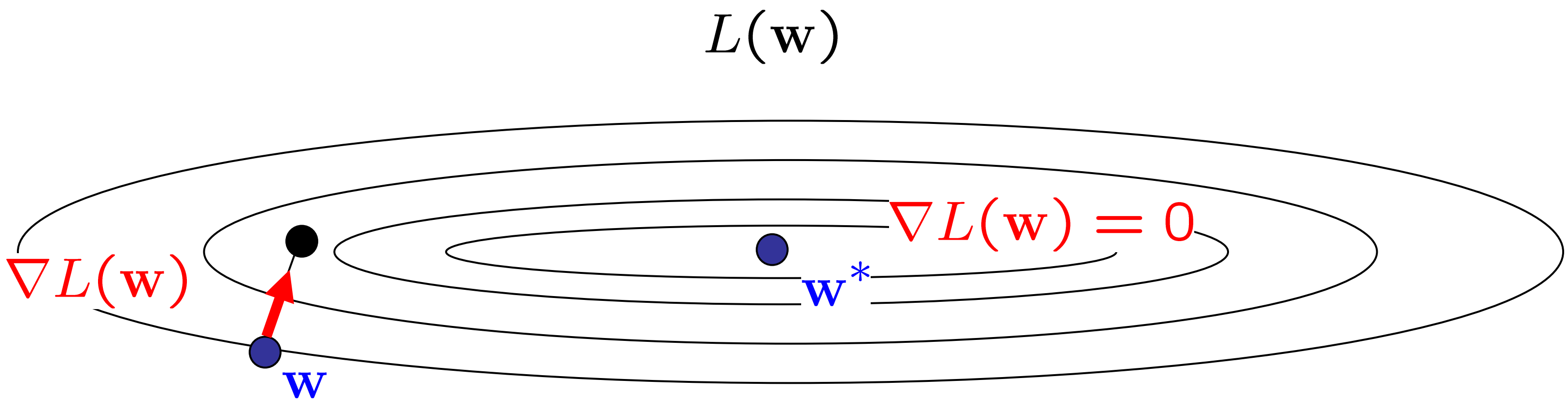
$$\log\left(\left(\frac{e^x}{e^x+e^y}\right)^3 \times \frac{e^y}{e^x+e^y}\right)$$





Convex Optimization

- The maxent objective is an unconstrained convex problem



- One optimal value*, gradients point the way

Gradients

$$L(\mathbf{w}) = \sum_i \left(\mathbf{w}^\top \mathbf{f}(\mathbf{x}_i, \mathbf{y}_i^*) - \log \sum_{\mathbf{y}} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{x}_i, \mathbf{y})) \right)$$

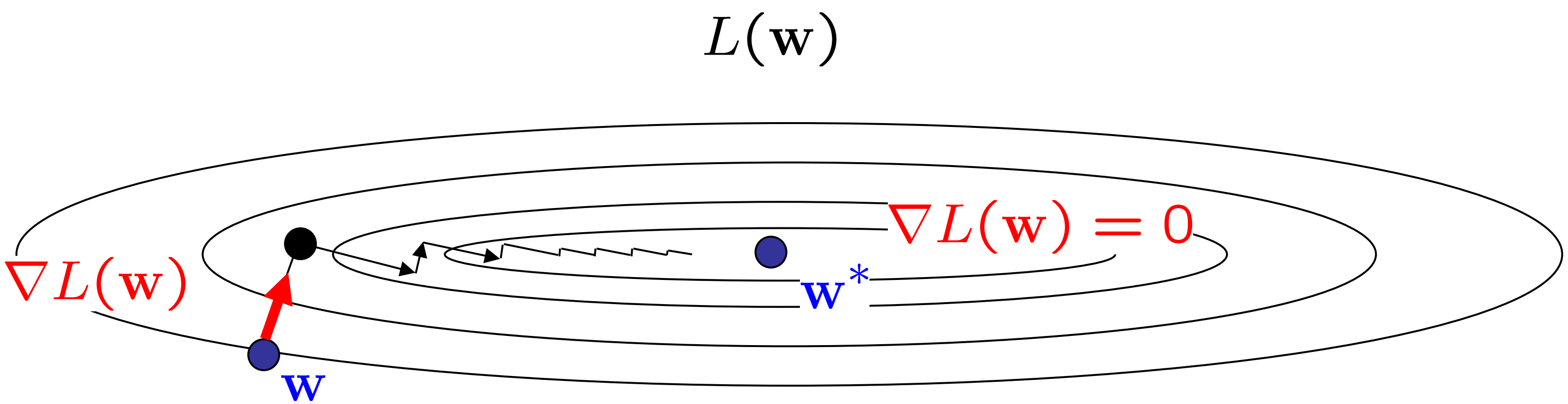
$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \sum_i \left(\mathbf{f}(\mathbf{x}_i, \mathbf{y}_i^*) - \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}_i) \mathbf{f}(\mathbf{x}_i, \mathbf{y}) \right)$$

Count of features under target labels

Expected count of features under model predicted label distribution

Gradient Ascent

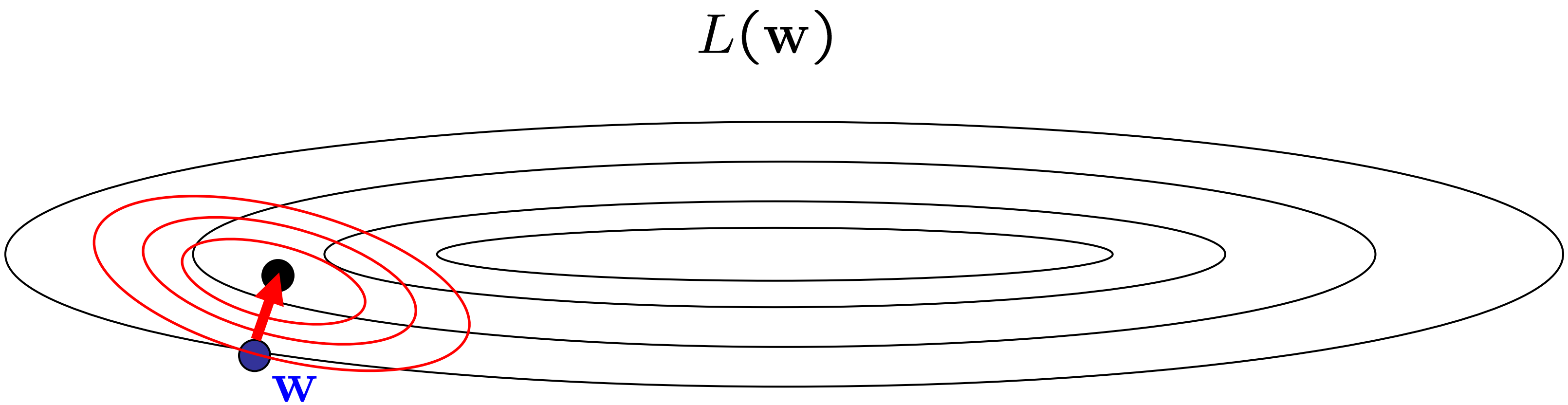
- The maxent objective is an unconstrained optimization problem



- Gradient Ascent
  - Basic idea: move uphill from current guess
  - Gradient ascent / descent follows the gradient incrementally
  - At local optimum, derivative vector is zero
  - Will converge if step sizes are small enough, but not efficient
  - All we need is to be able to evaluate the function and its derivative

(Quasi)-Newton Methods

- 2^{nd}-Order methods: repeatedly create a quadratic approximation and solve it



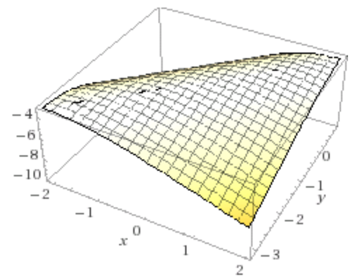
$$L(\mathbf{w}_0) + \nabla L(\mathbf{w})^\top (\mathbf{w} - \mathbf{w}_0) + (\mathbf{w} - \mathbf{w}_0)^\top \nabla^2 L(\mathbf{w}) (\mathbf{w} - \mathbf{w}_0)$$

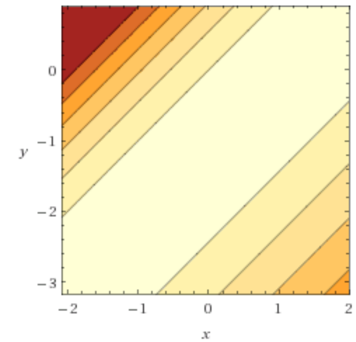
- E.g. LBFGS, which tracks derivative to approximate (inverse) Hessian

Regularization

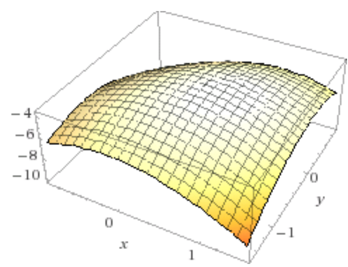
Regularization Methods

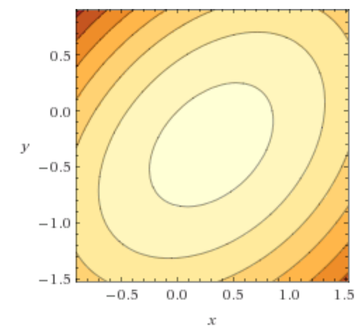
- Early stopping



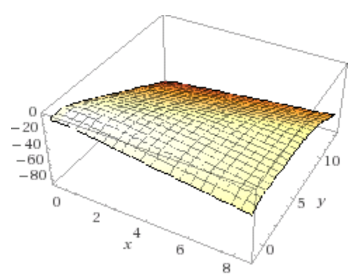


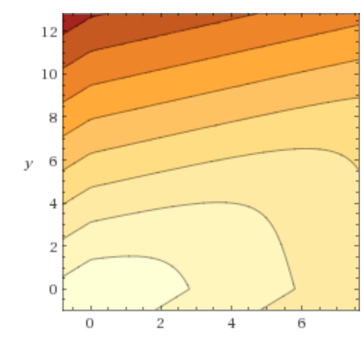
- L2: $L(w)-|w|_2^2$





- L1: $L(w)-|w|$





Regularization Effects

- Early stopping: don’t do this
- L2: weights stay small but non-zero
- L1: many weights driven to zero
  - Good for sparsity
  - Usually bad for accuracy for NLP

Scaling

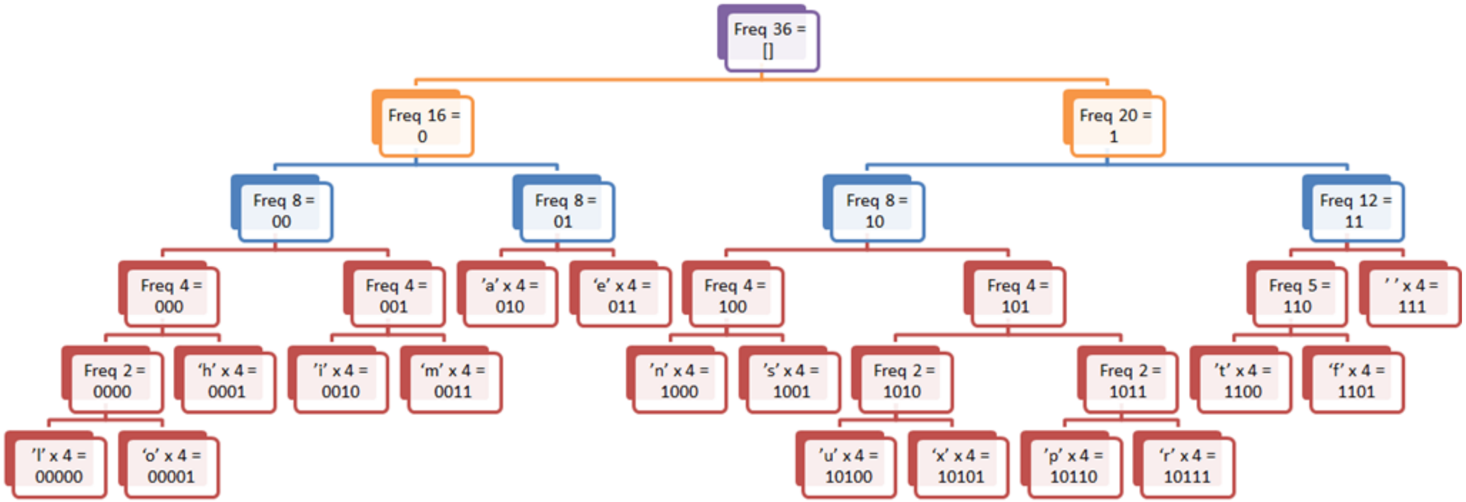
Why is Scaling Hard?

$$L(\mathbf{w}) = \sum_i \left(\mathbf{w}^\top \mathbf{f}(\mathbf{x}_i, \mathbf{y}_i^*) - \log \sum_{\mathbf{y}} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{x}_i, \mathbf{y})) \right)$$

- Big normalization terms
- Lots of data points

Hierarchical Prediction

- Hierarchical prediction / softmax [Mikolov et al 2013]



- Noise-Contrastive Estimation [Mnih, 2013]

- Self-Normalization [Devlin, 2014]

Image: ayende.com

Stochastic Gradient

- View the gradient as an average over data points

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{N} \sum_i \left(\mathbf{f}(\mathbf{x}_i, \mathbf{y}_i^*) - \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}_i) \mathbf{f}(\mathbf{x}_i, \mathbf{y}) \right)$$

- Stochastic gradient: take a step each example (or mini-batch)

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} \approx \frac{1}{1} \left(\mathbf{f}(\mathbf{x}_i, \mathbf{y}_i^*) - \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}_i) \mathbf{f}(\mathbf{x}_i, \mathbf{y}) \right)$$

- Substantial improvements exist, e.g. AdaGrad (Duchi, 11)

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

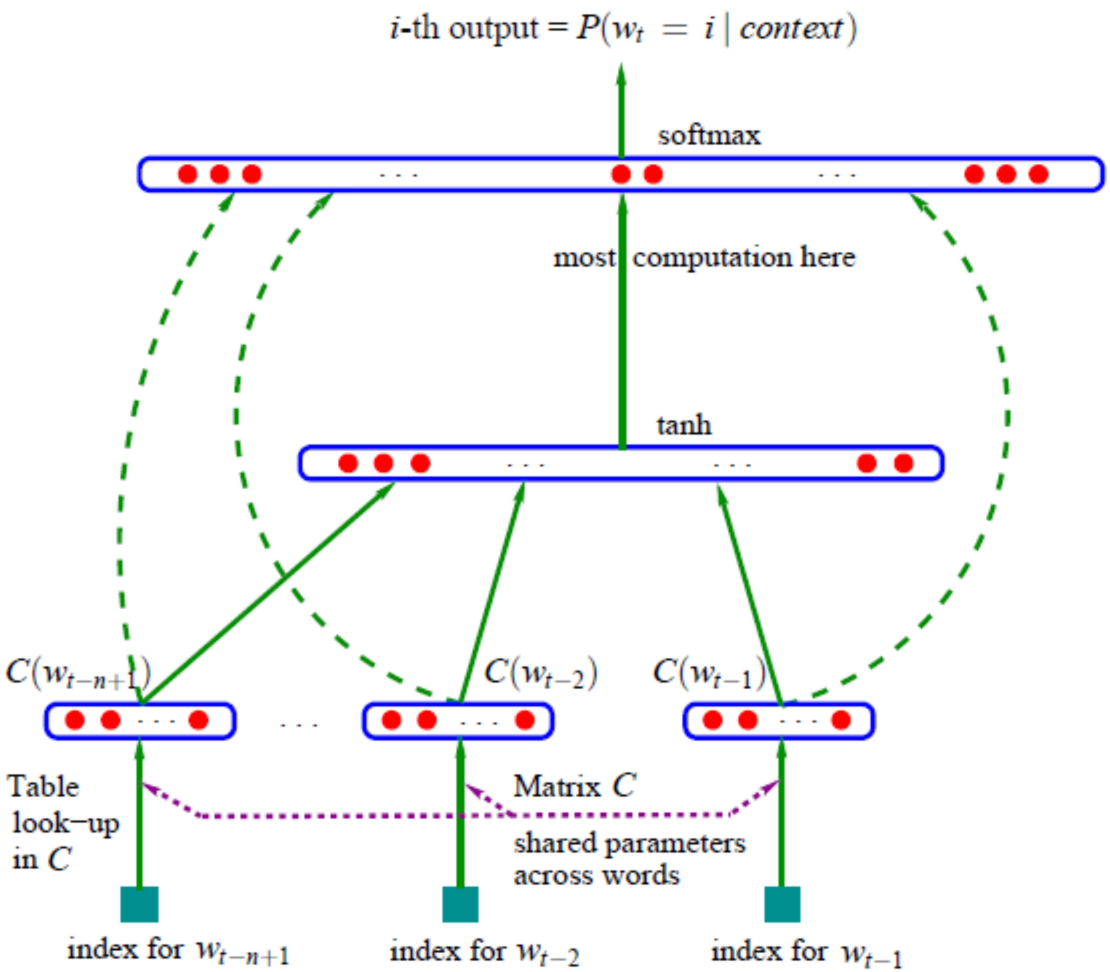


Image: (Bengio et al, 03)

Neural vs Maxent

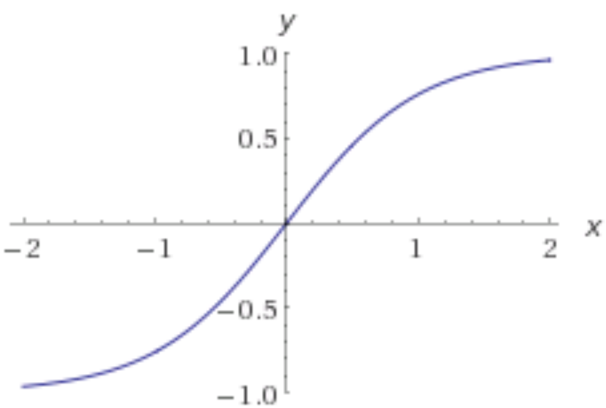
- Maxent LM

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

- Simple Neural LM

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

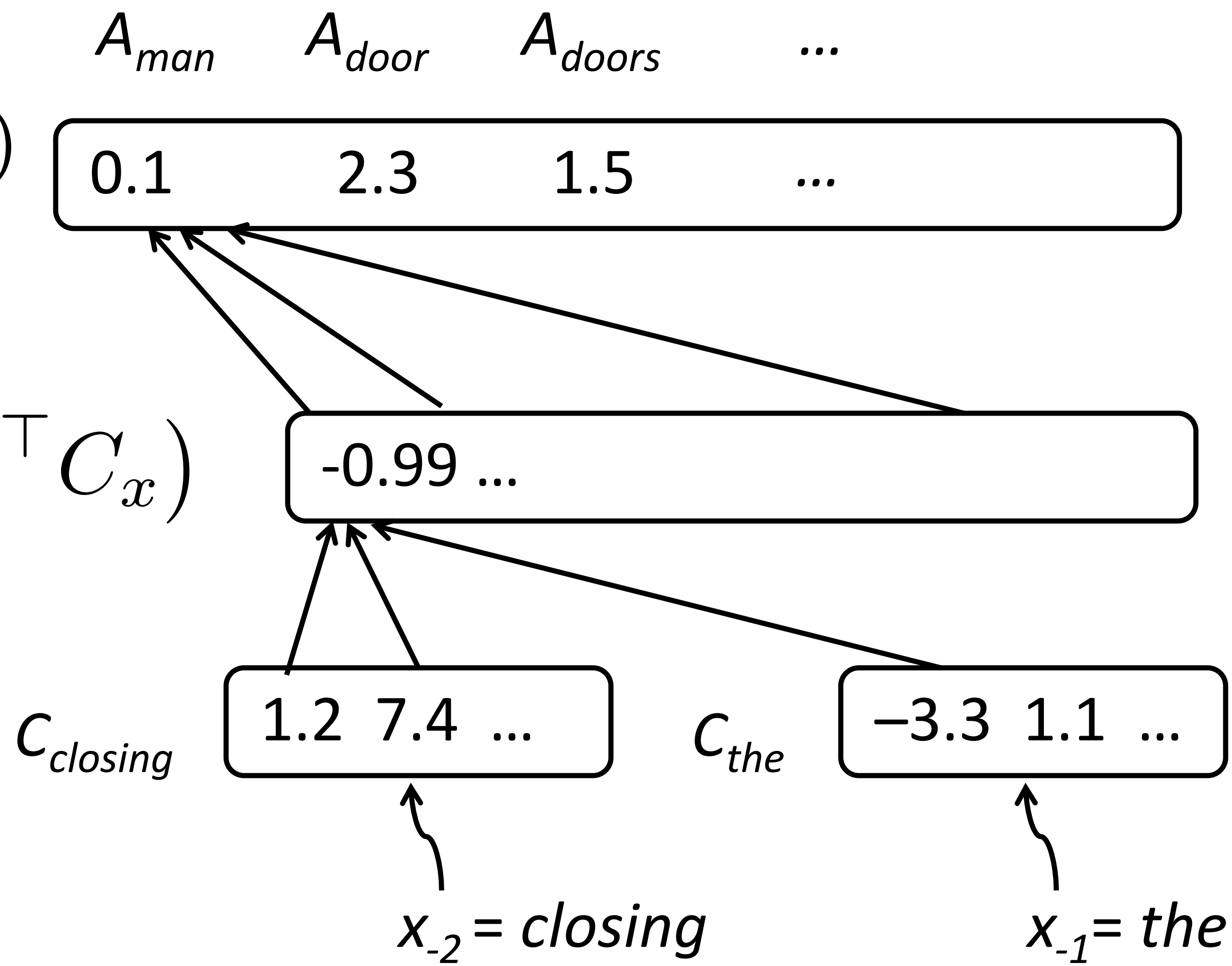
σ nonlinear, e.g. tanh



Neural LM Example

$$P(y|x) \propto e^{(A^\top h)}$$

$$h = \sigma\left(B^\top C_x\right)$$



Neural LMs

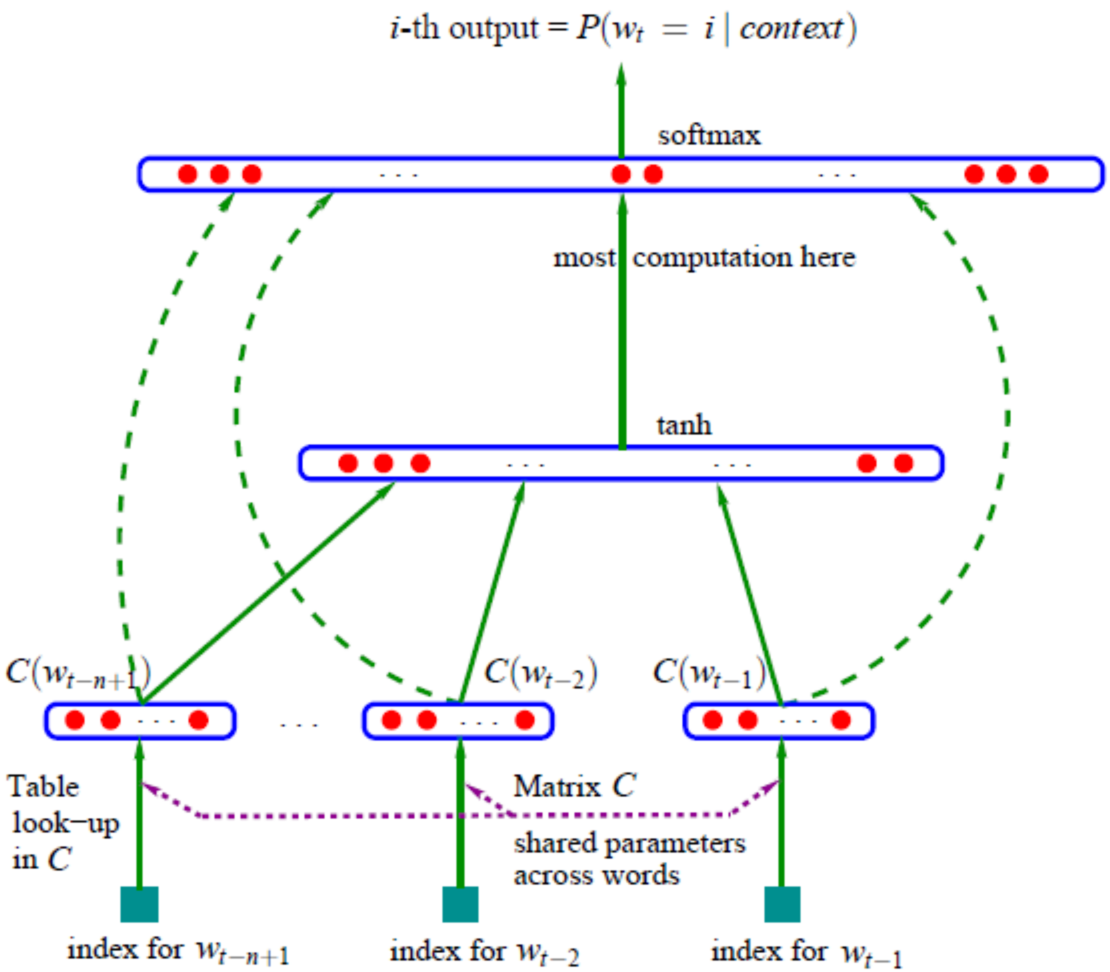
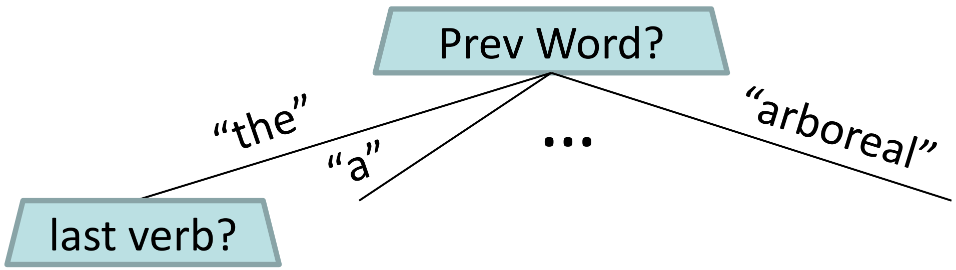


Image: (Bengio et al, 03)

Decision Trees / Forests

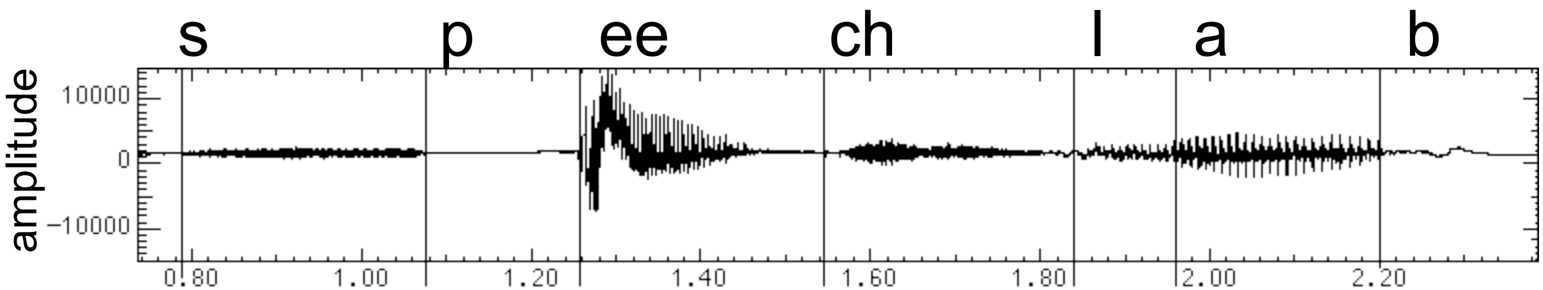


- Decision trees?
  - Good for non-linear decision problems
  - Random forests can improve further [Xu and Jelinek, 2004]
  - Paths to leaves basically learn conjunctions
  - General contrast between DTs and linear models

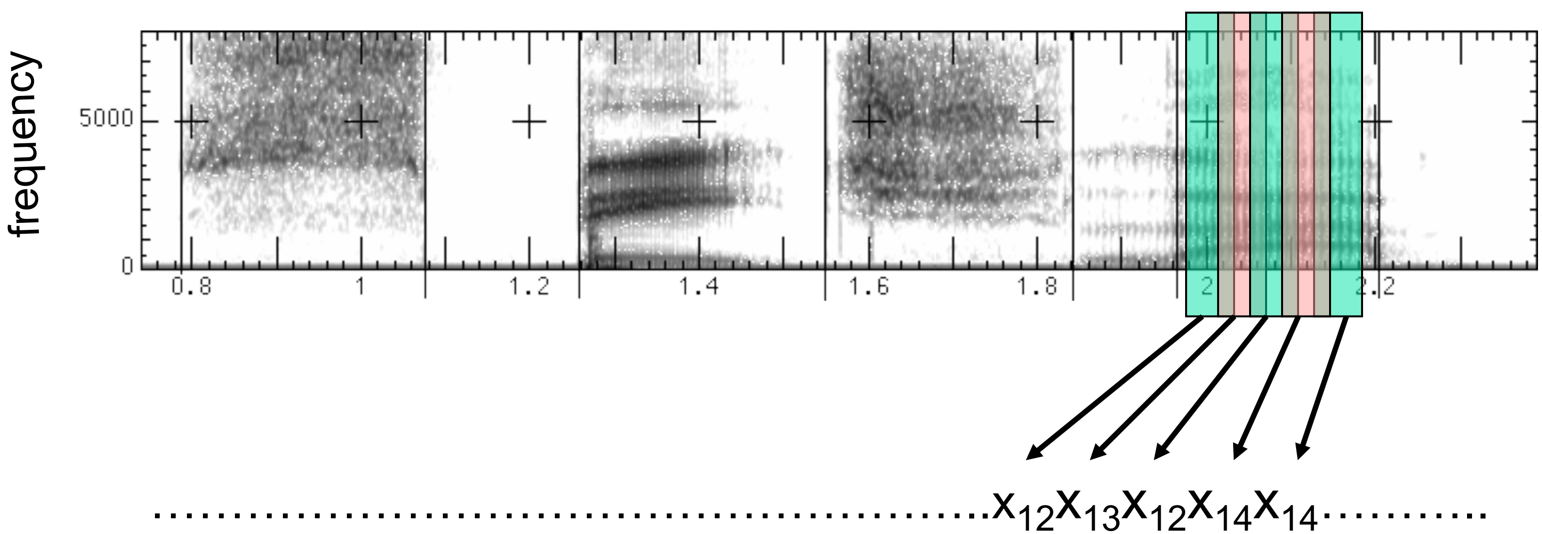
Speech Signals

Speech in a Slide

- Frequency gives pitch; amplitude gives volume

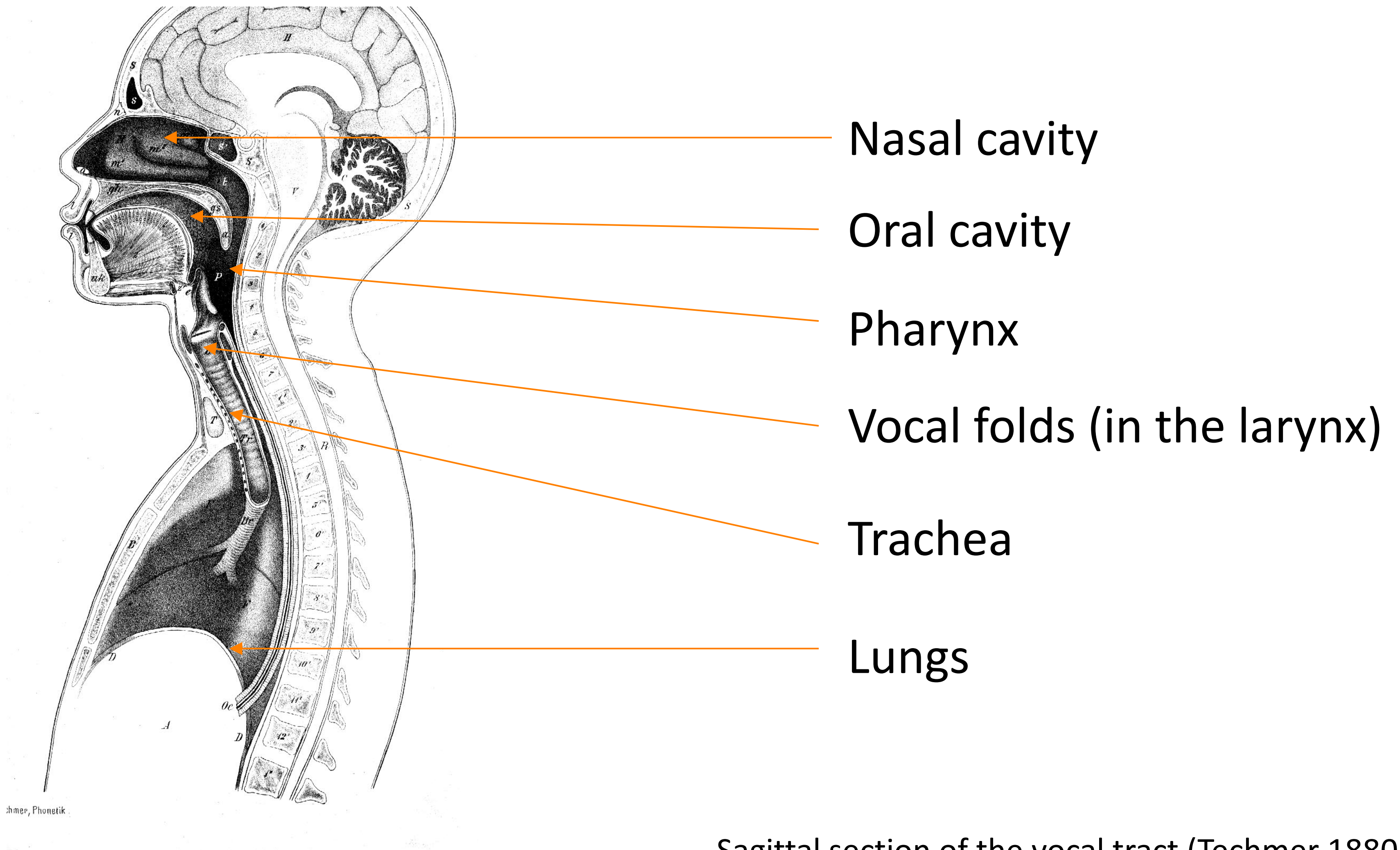


- Frequencies at each time slice processed into observation vectors



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		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

Place

Places of Articulation

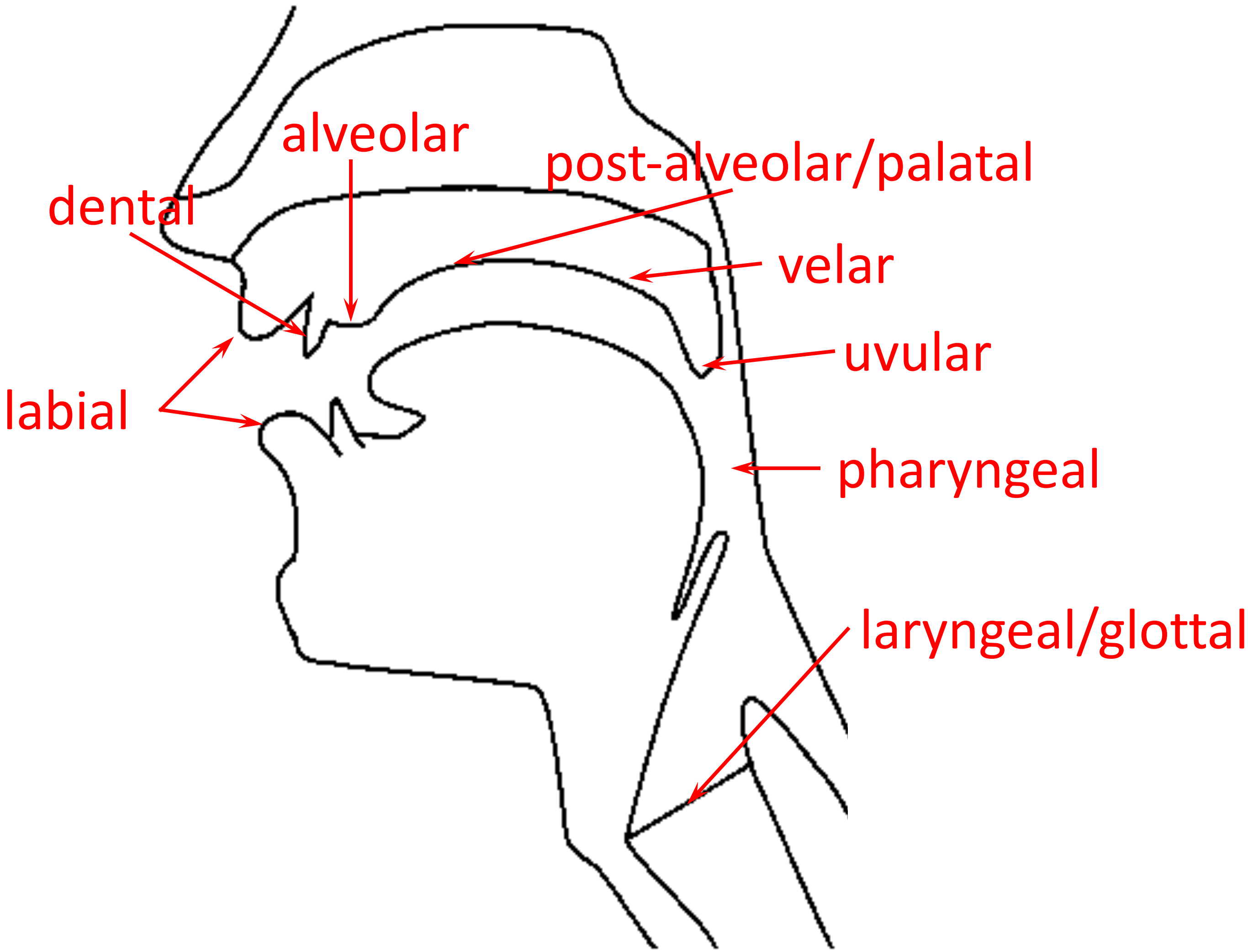
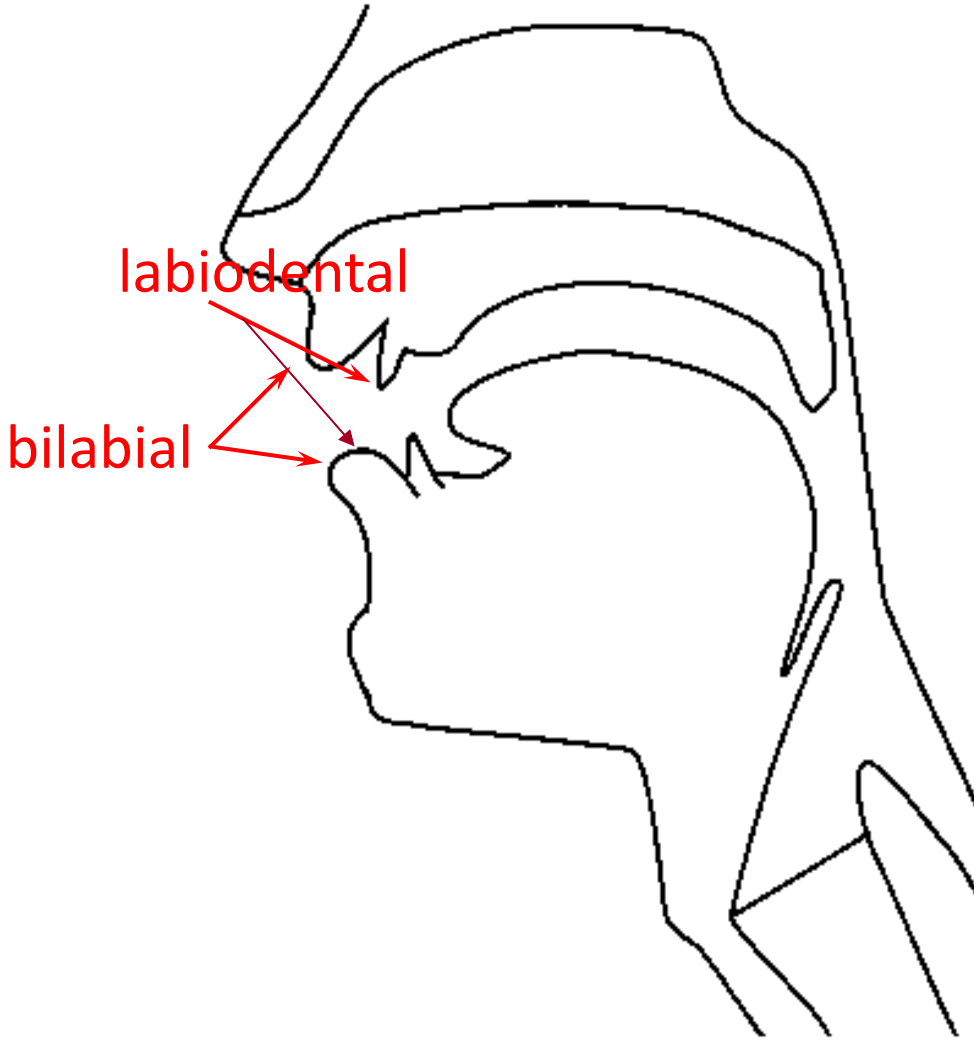


Figure thanks to Jennifer Venditti

Labial place



Bilabial:

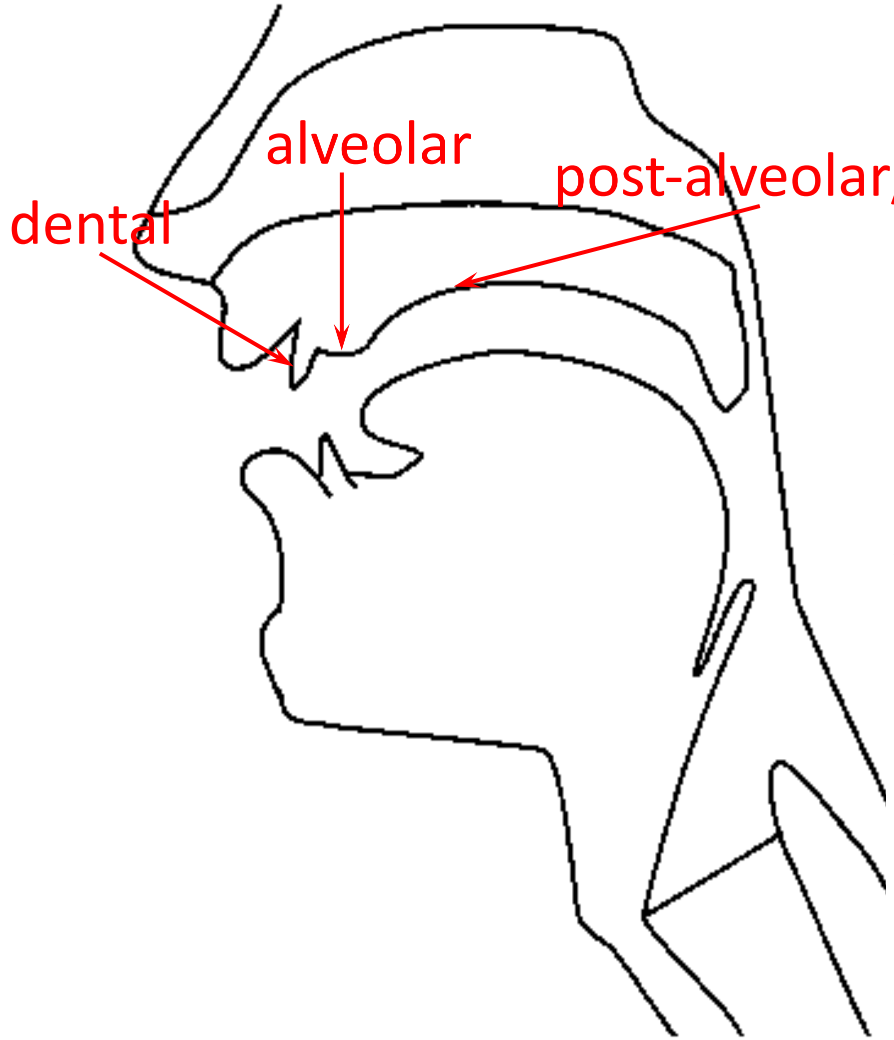
p, b, m

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

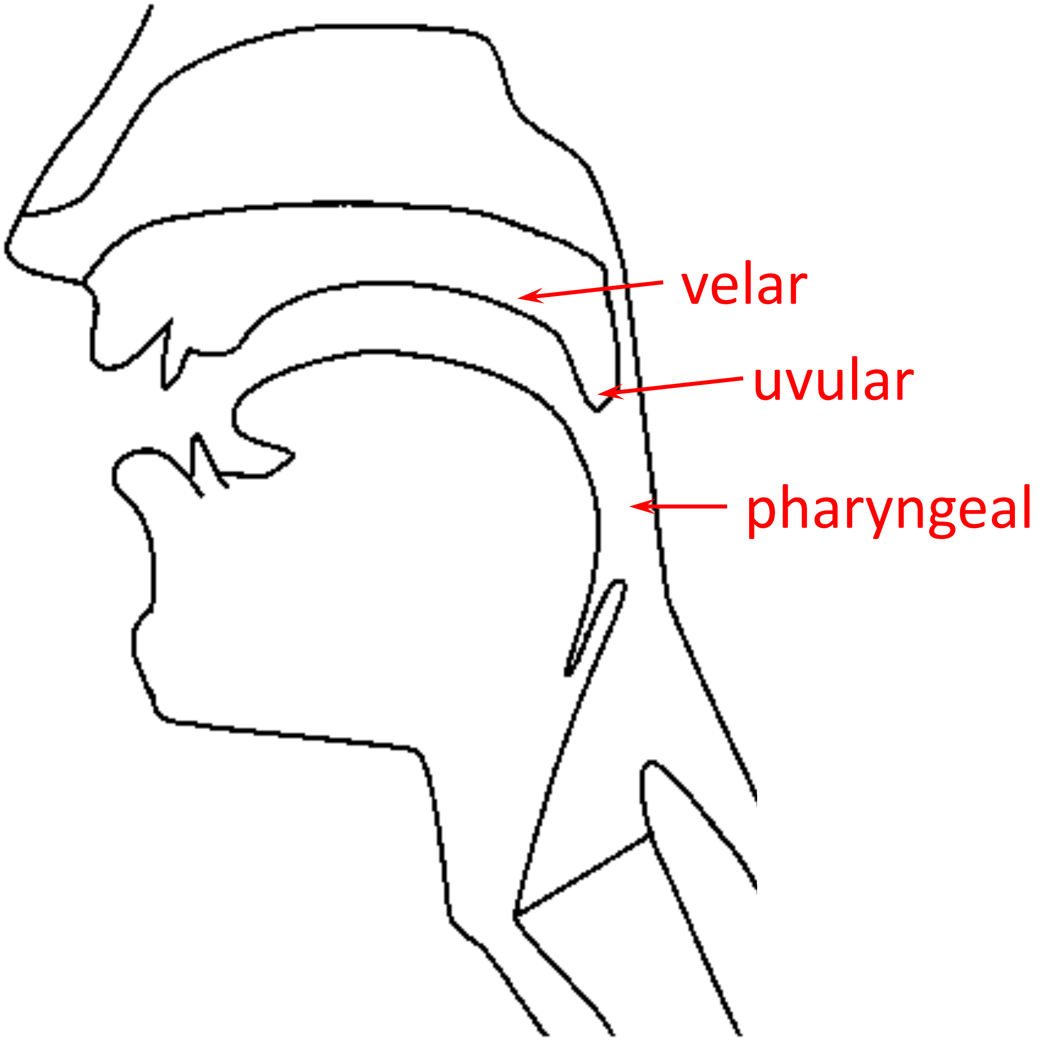


Figure thanks to Jennifer Venditti

Space of Phonemes

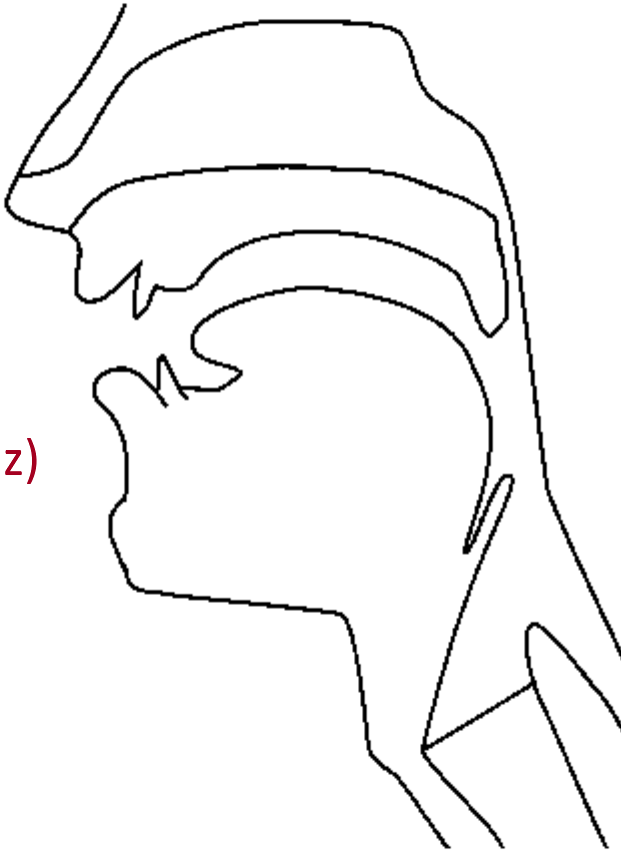
	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		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

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)



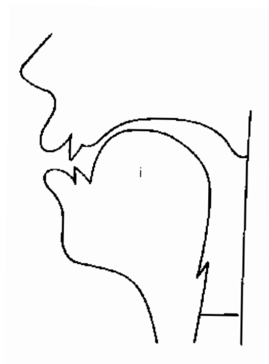
Space of Phonemes

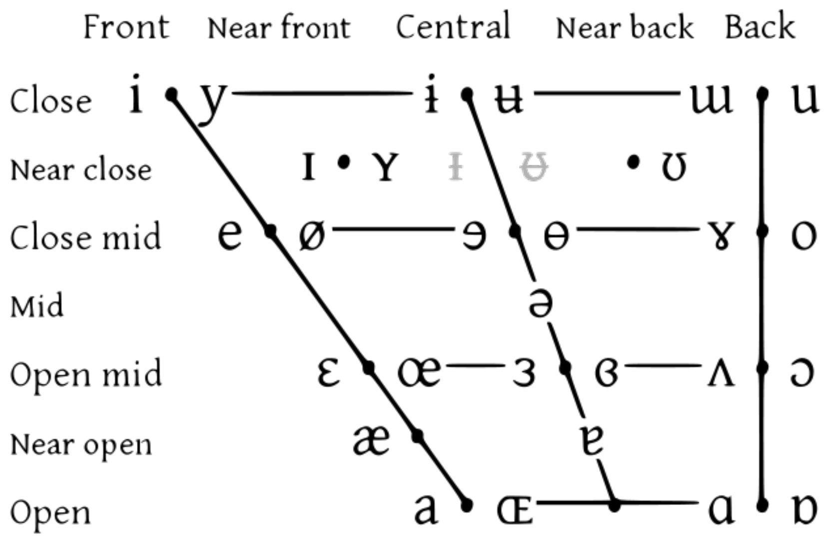
	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant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

- Standard international phonetic alphabet (IPA) chart of consonants

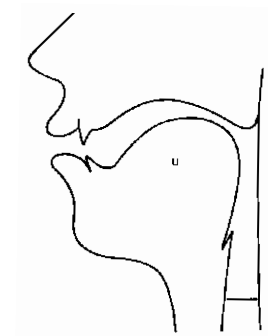
Vowels

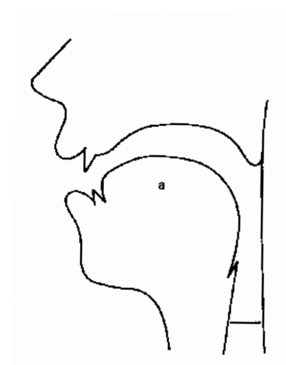
Vowel Space





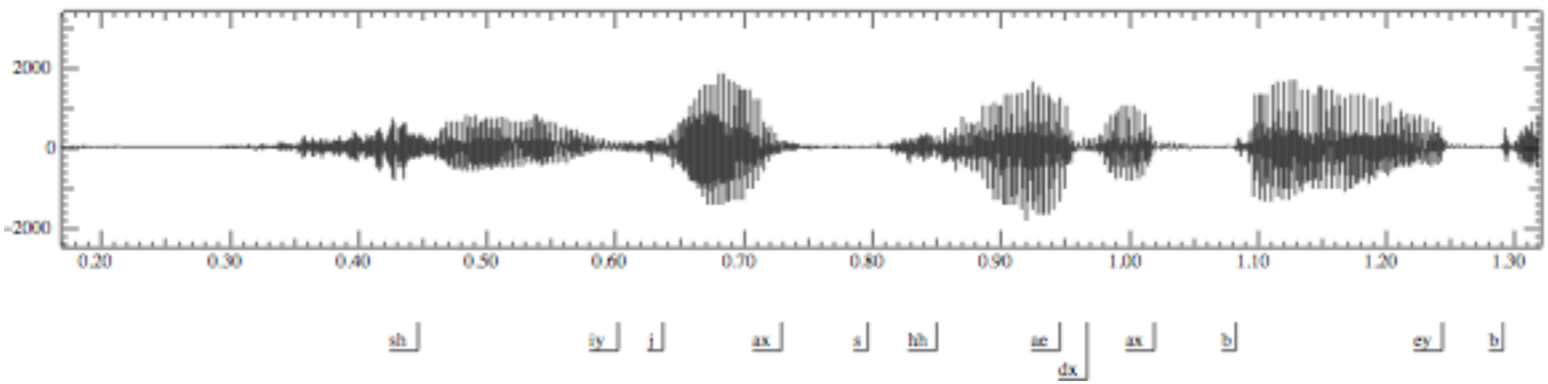
Vowels at right & left of bullets are rounded & unrounded.





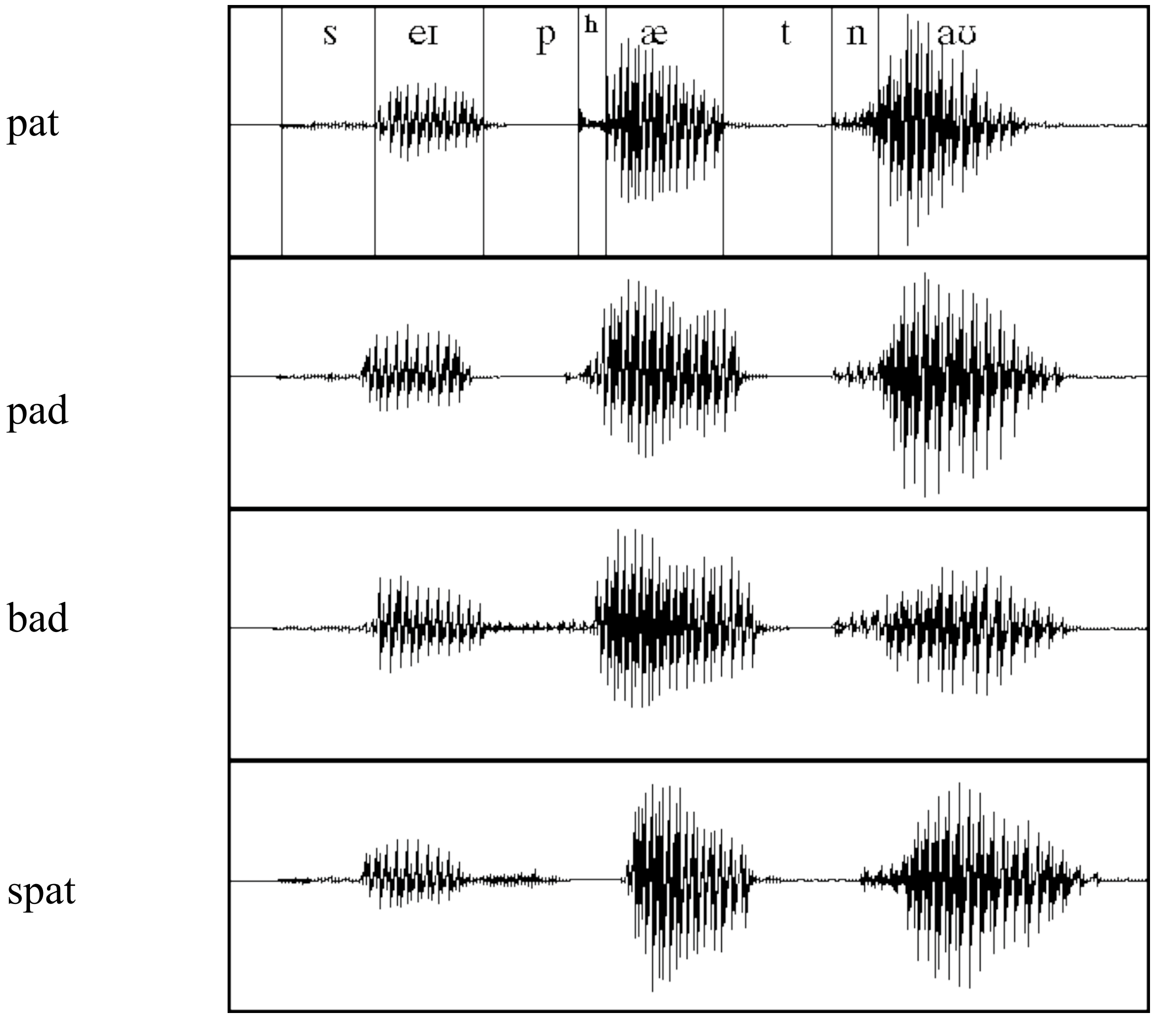
Acoustics

“She just had a baby”



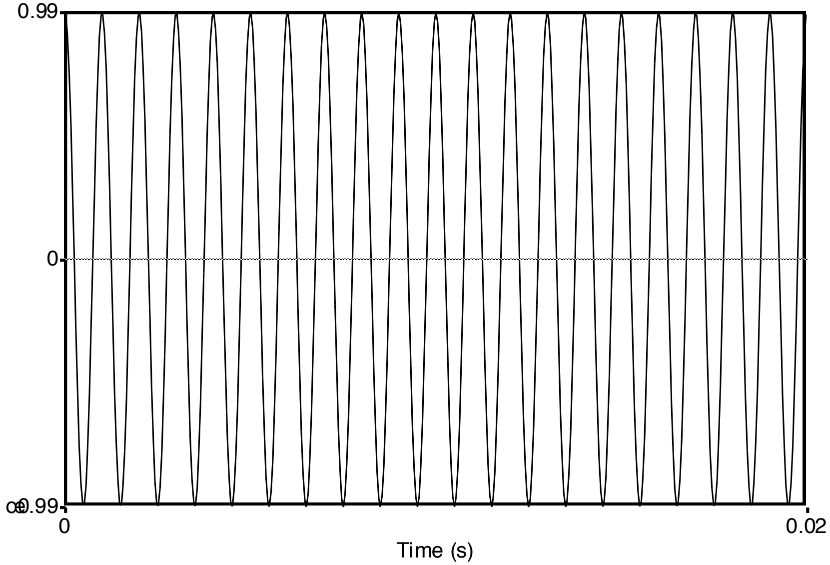
- What can we learn from a wavefile?
  - No gaps between words (!)
  - Vowels are voiced, long, loud
  - Length in time = length in space in waveform picture
  - Voicing: regular peaks in amplitude
  - When stops closed: no peaks, silence
  - Peaks = voicing: .46 to .58 (vowel [iy], from second .65 to .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



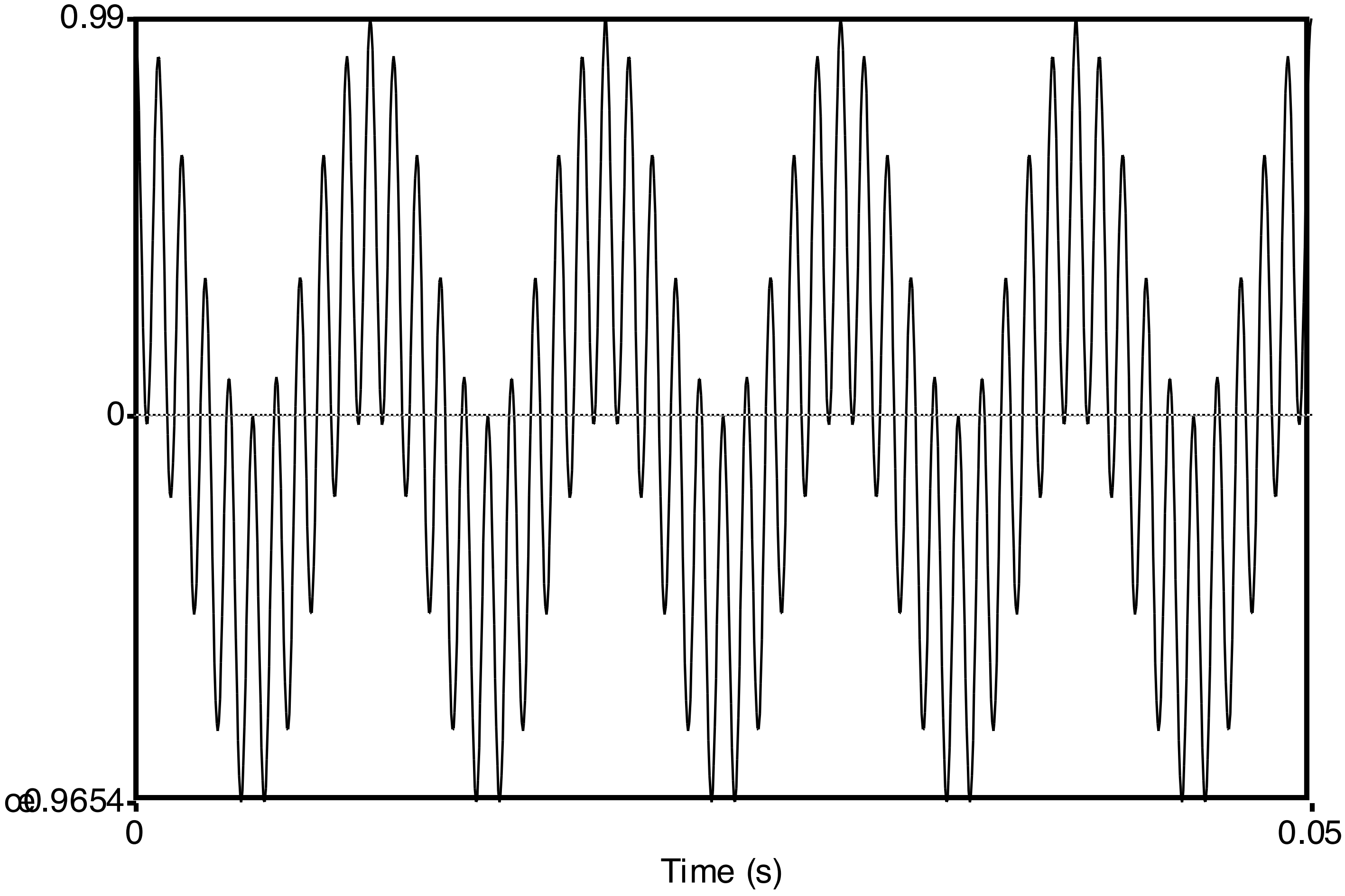
Example from Ladefoged

Simple Periodic Waves of Sound



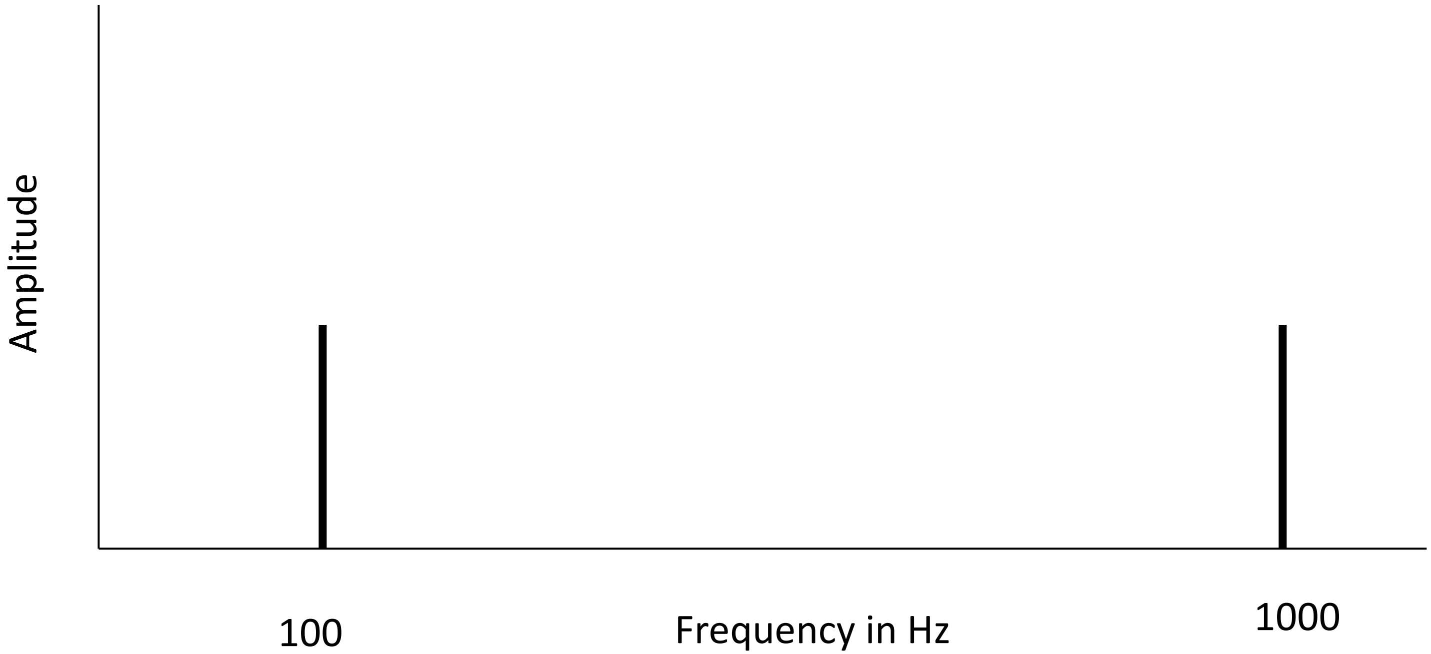
- Y axis: Amplitude = amount of air pressure at that point in time
  - Zero is normal air pressure, negative is rarefaction
- X axis: Time.
- Frequency = number of cycles per second.
- 20 cycles in .02 seconds = 1000 cycles/second = 1000 Hz

Complex Waves: 100Hz+1000Hz

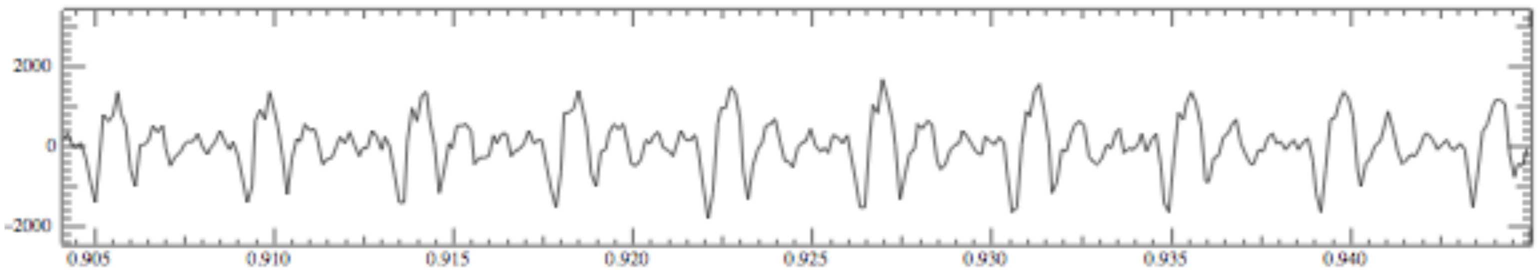


Spectrum

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

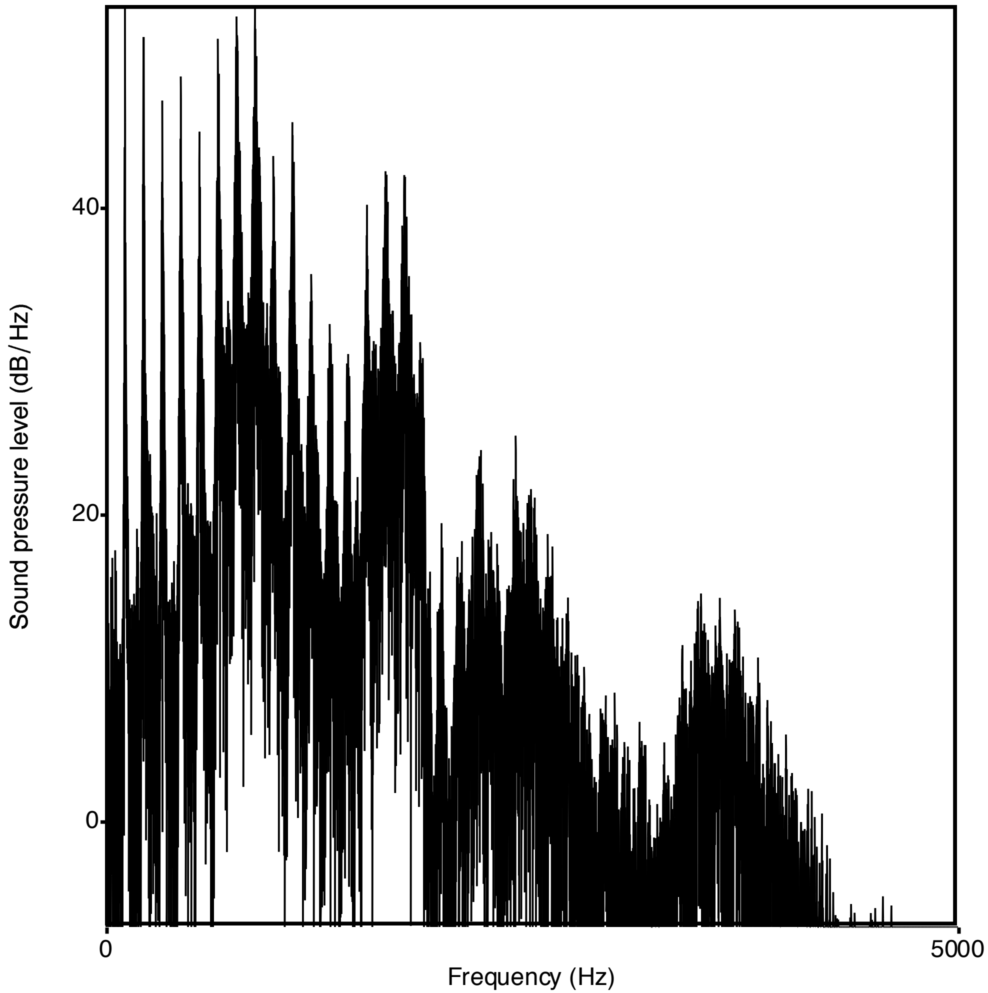


Part of [ae] waveform from "had"



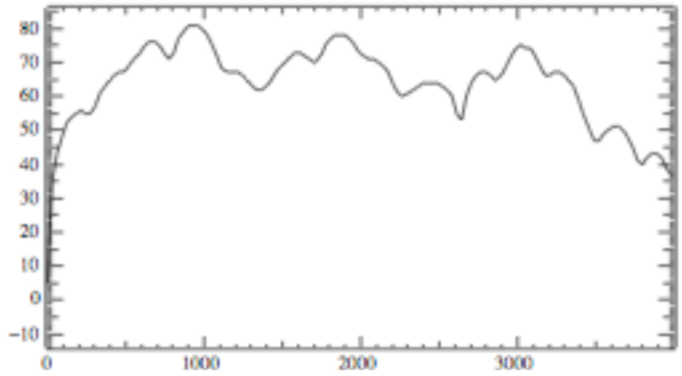
- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves

Spectrum of an Actual Soundwave



Back to Spectra

- Spectrum represents these freq components
- Computed by Fourier transform, algorithm which separates out each frequency component of wave.

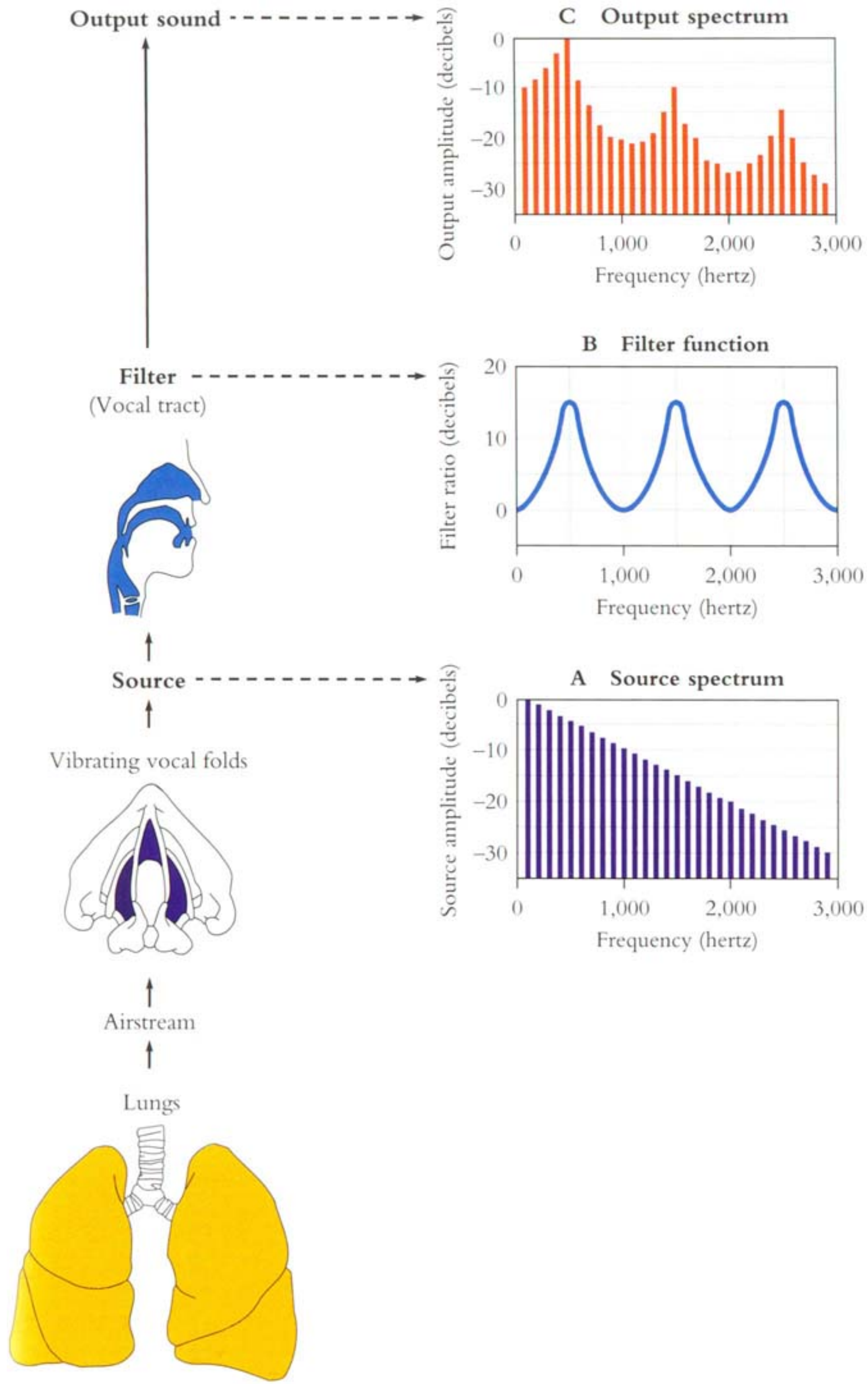


- x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

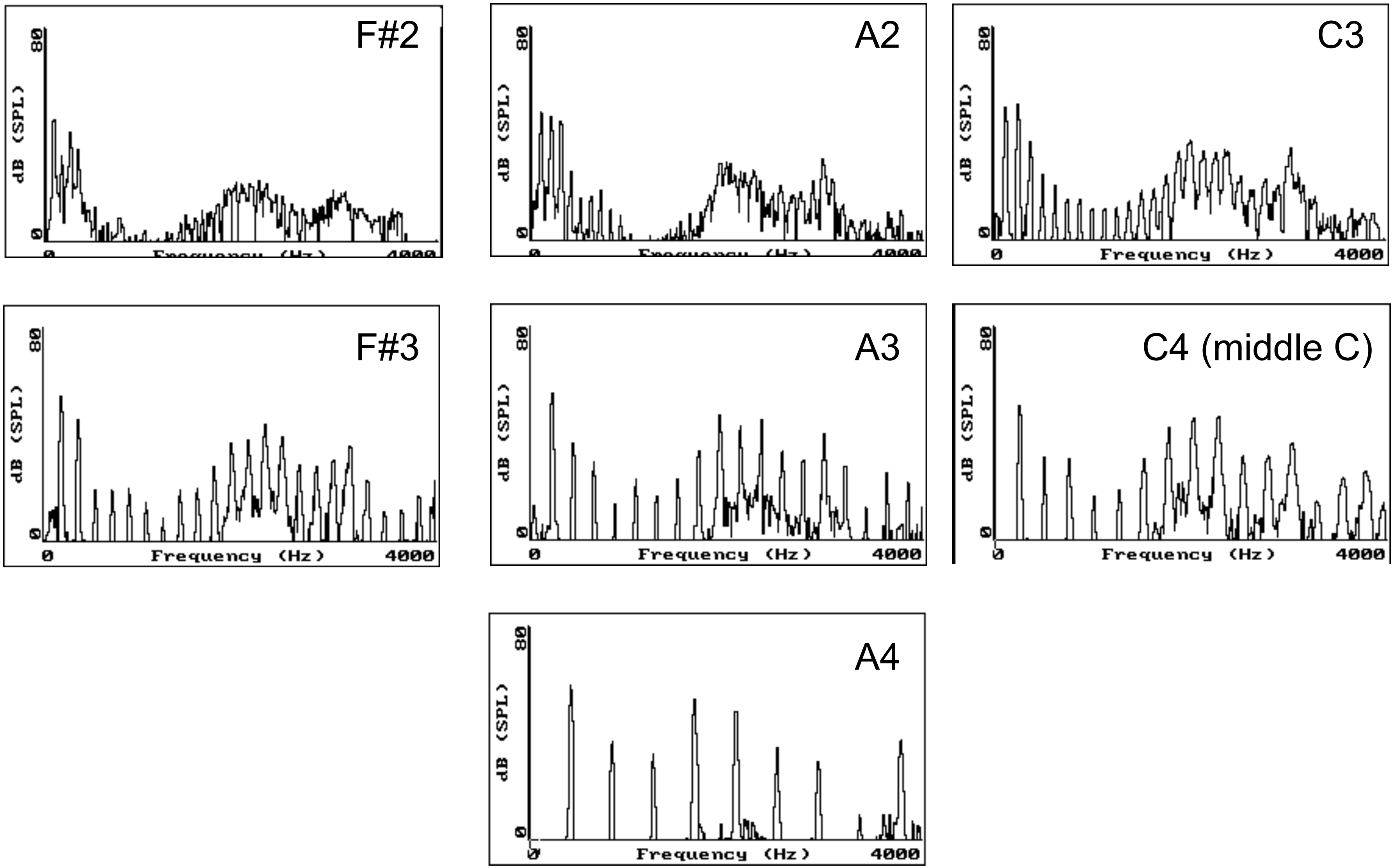
Source / Channel

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



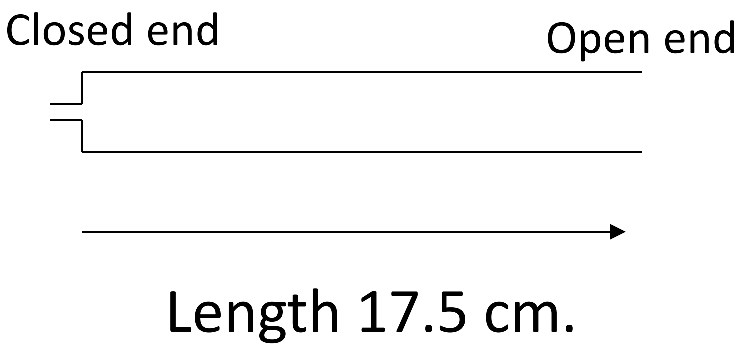
Vowel [i] at increasing pitches

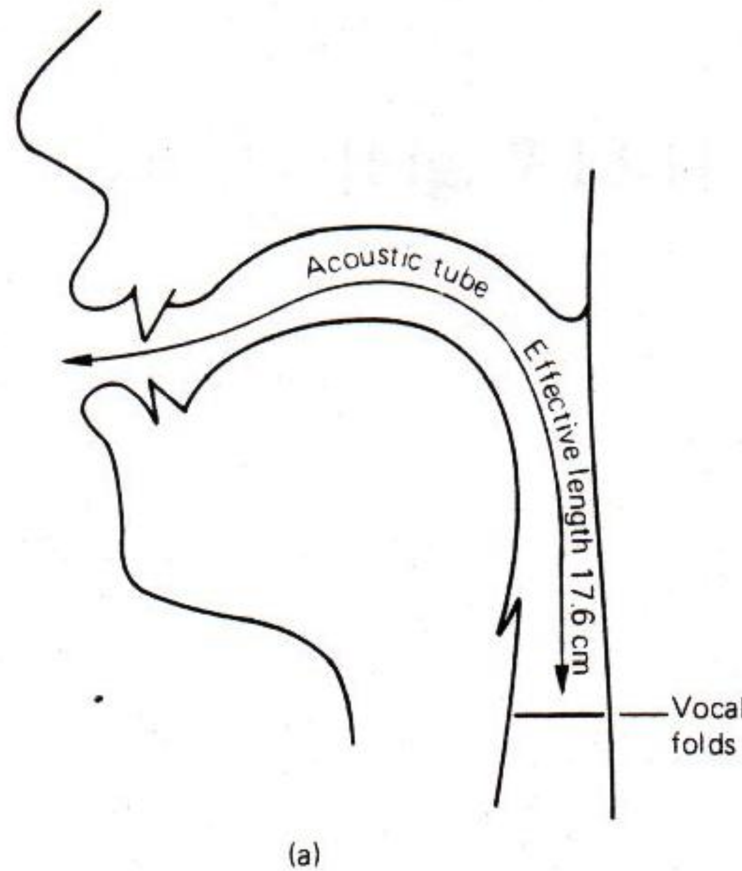


Figures from Ratree Wayland

Resonances of the Vocal Tract

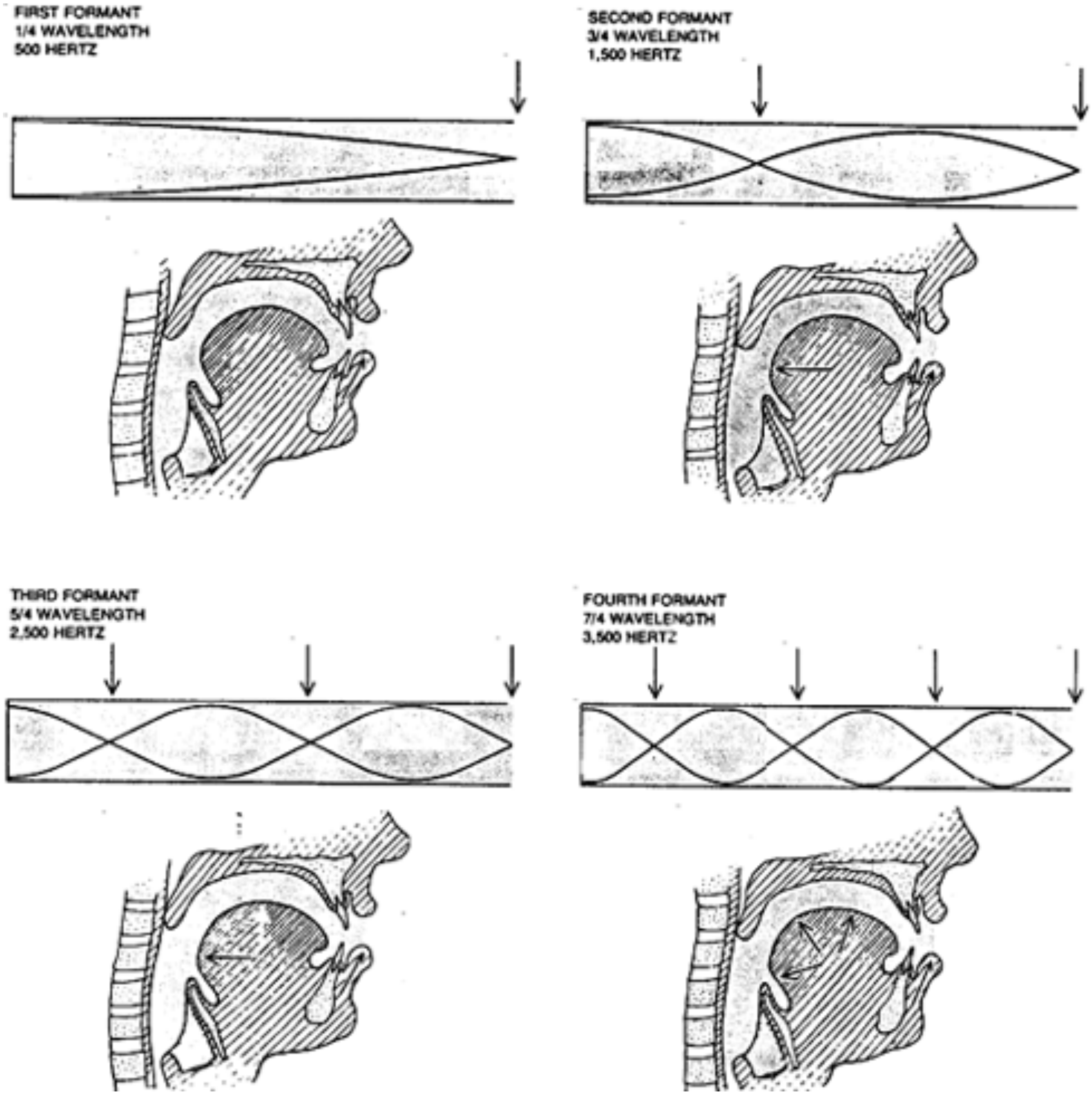
- The human vocal tract as an open tube:





- 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.

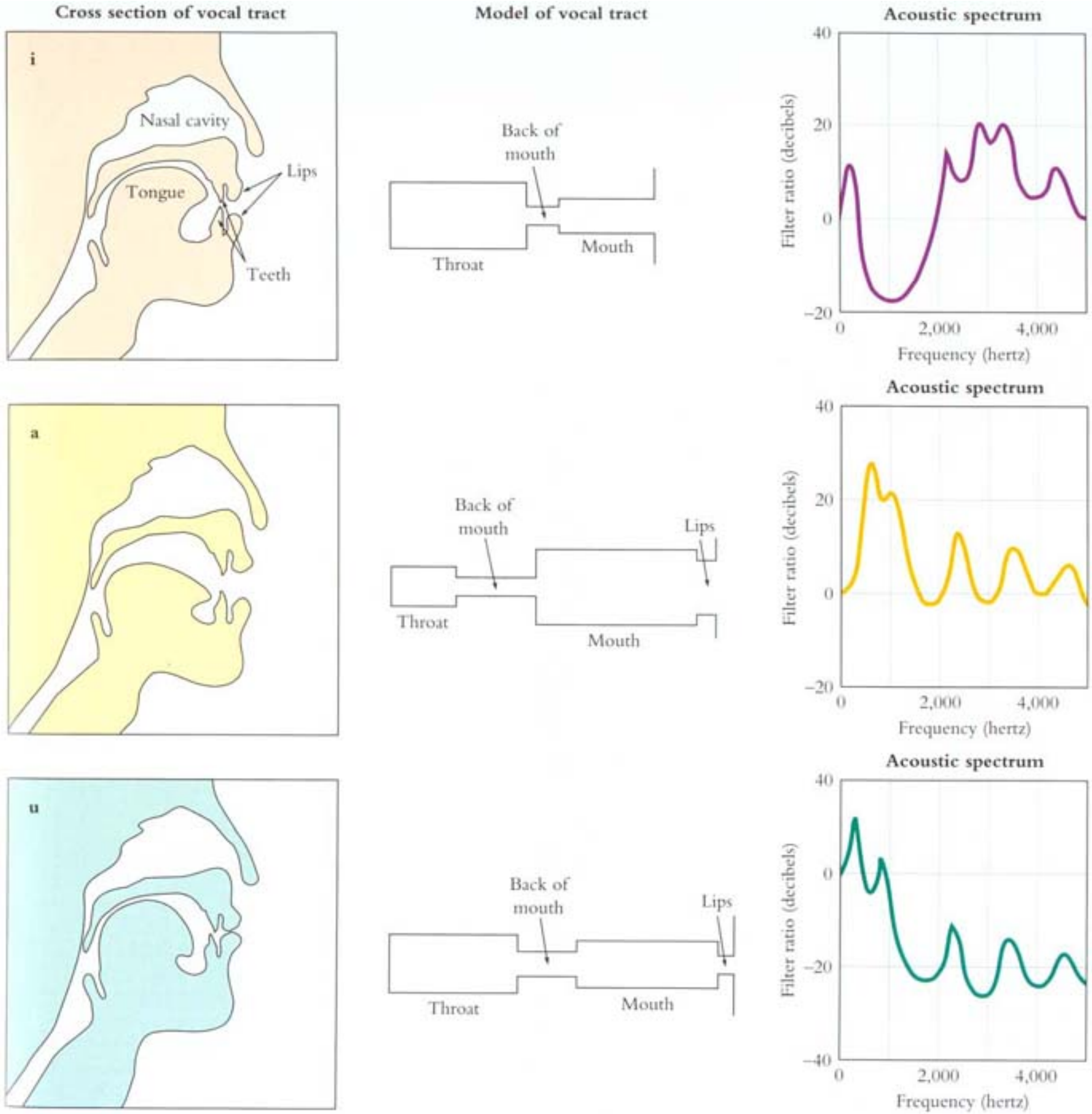
Figure from W. Barry



From Sundberg

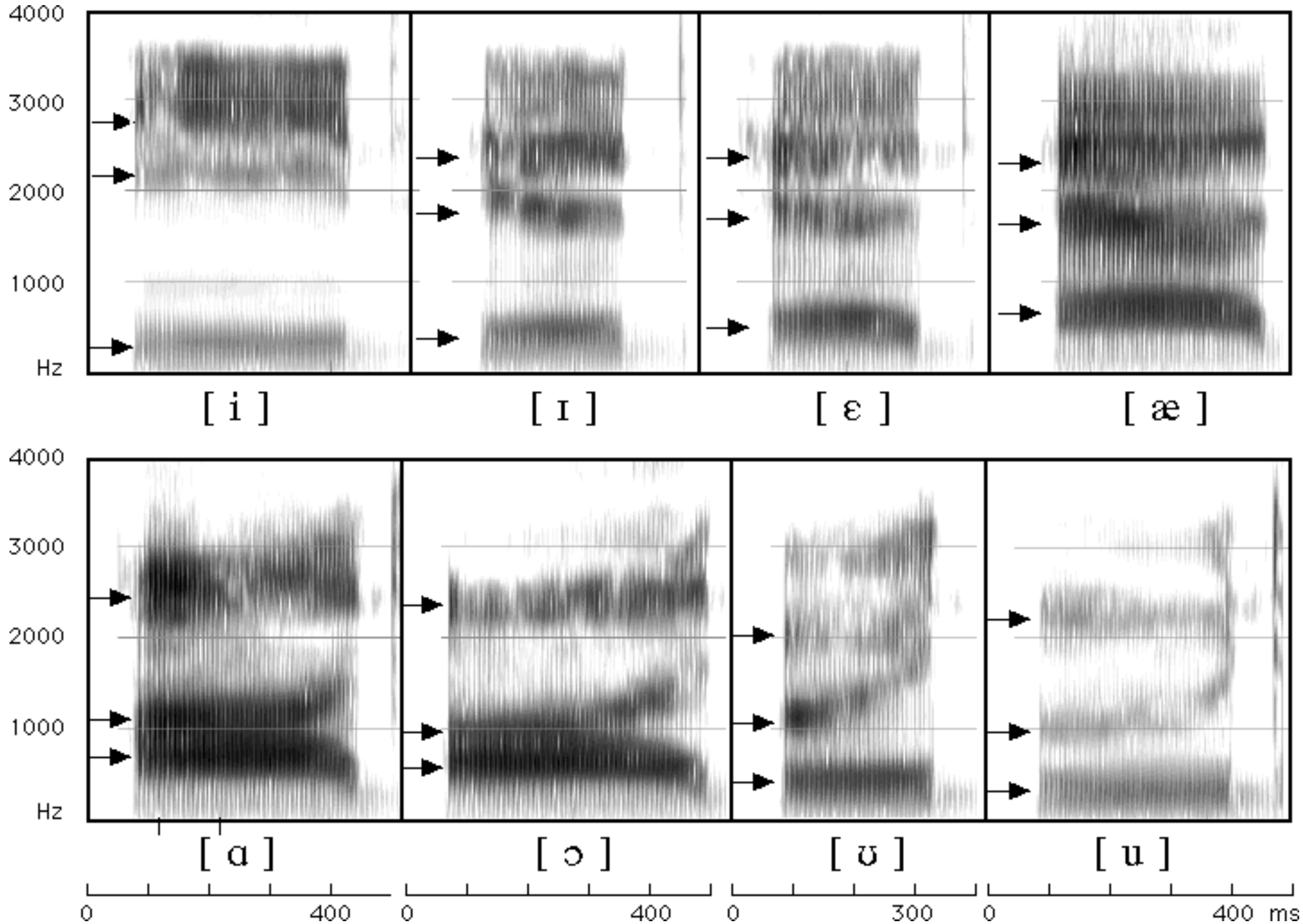
Computing the 3 Formants of Schwa

- Let the length of the tube be L
  - $F_1 = c/\lambda_1 = c/(4L) = 35{,}000/4*17.5 = 500\text{Hz}$
  - $F_2 = c/\lambda_2 = c/(4/3L) = 3c/4L = 3*35{,}000/4*17.5 = 1500\text{Hz}$
  - $F_3 = c/\lambda_3 = c/(4/5L) = 5c/4L = 5*35{,}000/4*17.5 = 2500\text{Hz}$
- So we expect a neutral vowel to have 3 resonances at 500, 1500, and 2500 Hz
- These vowel resonances are called formants

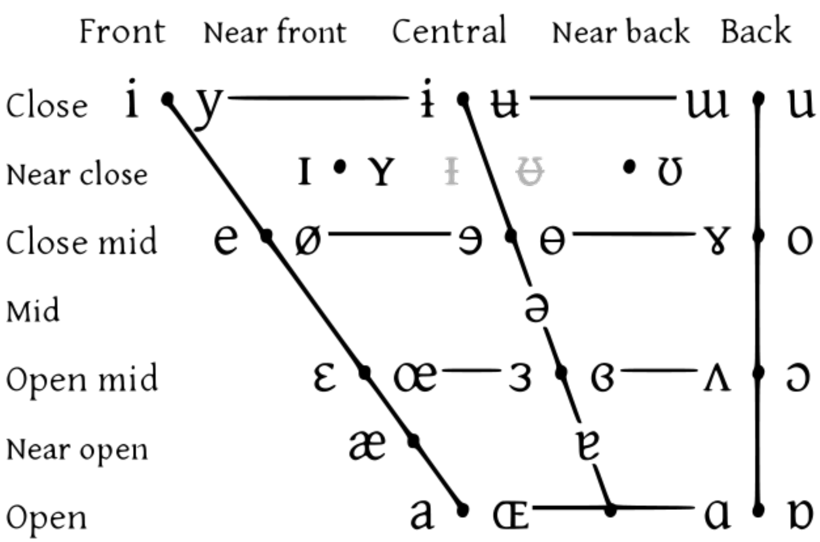


From
Mark
Liberman

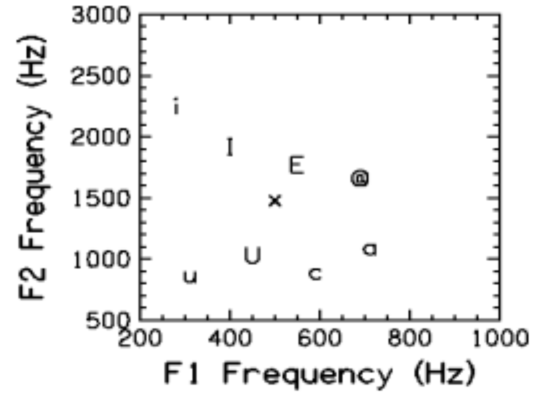
Seeing Formants: the Spectrogram

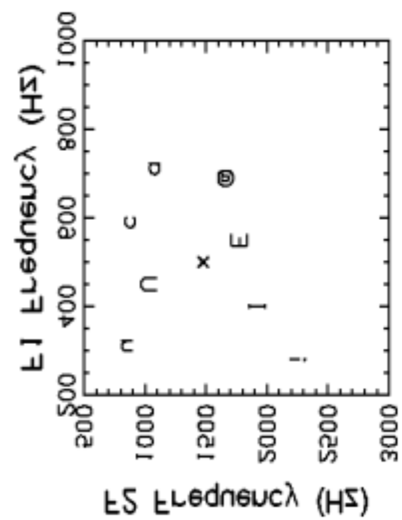


Vowel Space



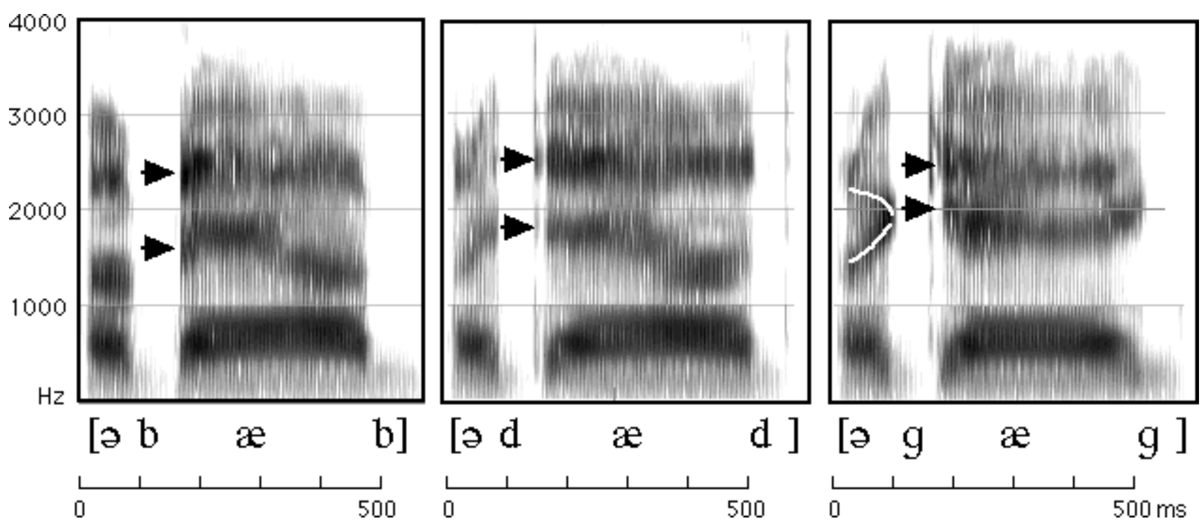
Vowels at right & left of bullets are rounded & unrounded.





Spectrograms

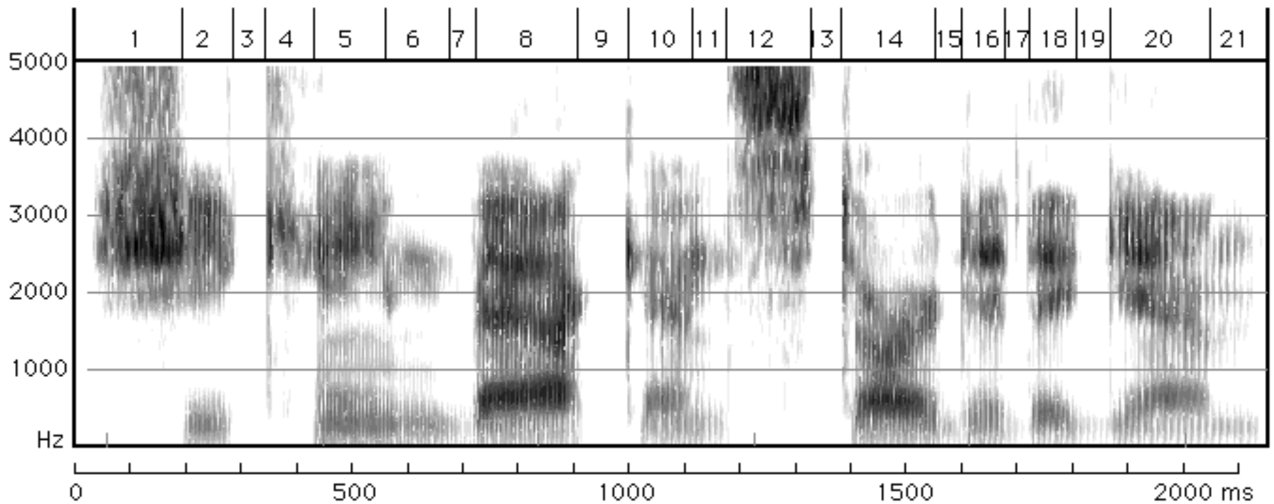
How to Read Spectrograms



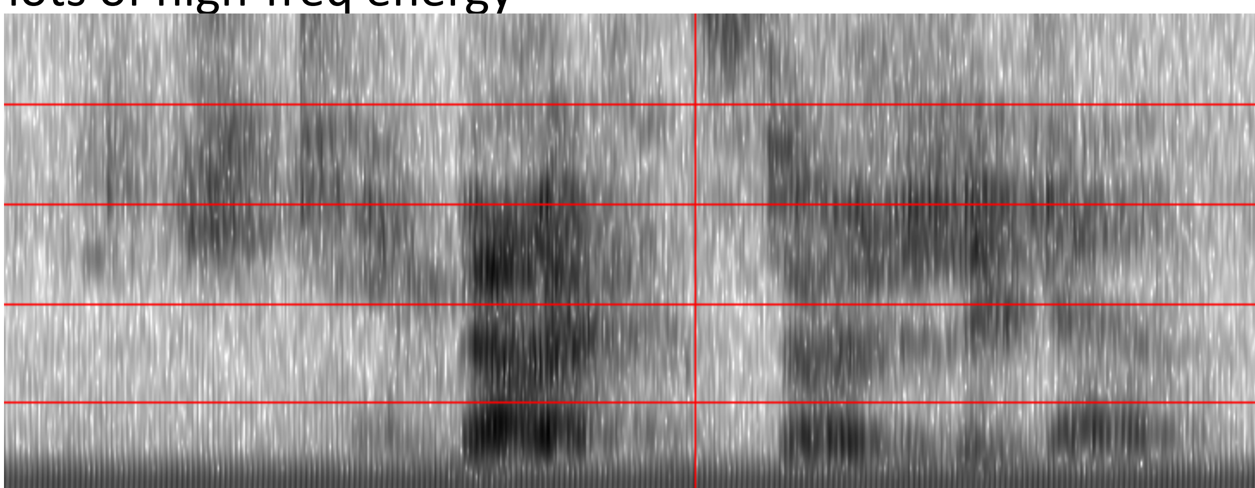
- [bab]: closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab"
- [dad]: first formant increases, but F2 and F3 slight fall
- [gag]: F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials

From Ladefoged "A Course in Phonetics"

“She came back and started again”



1. lots of high-freq energy

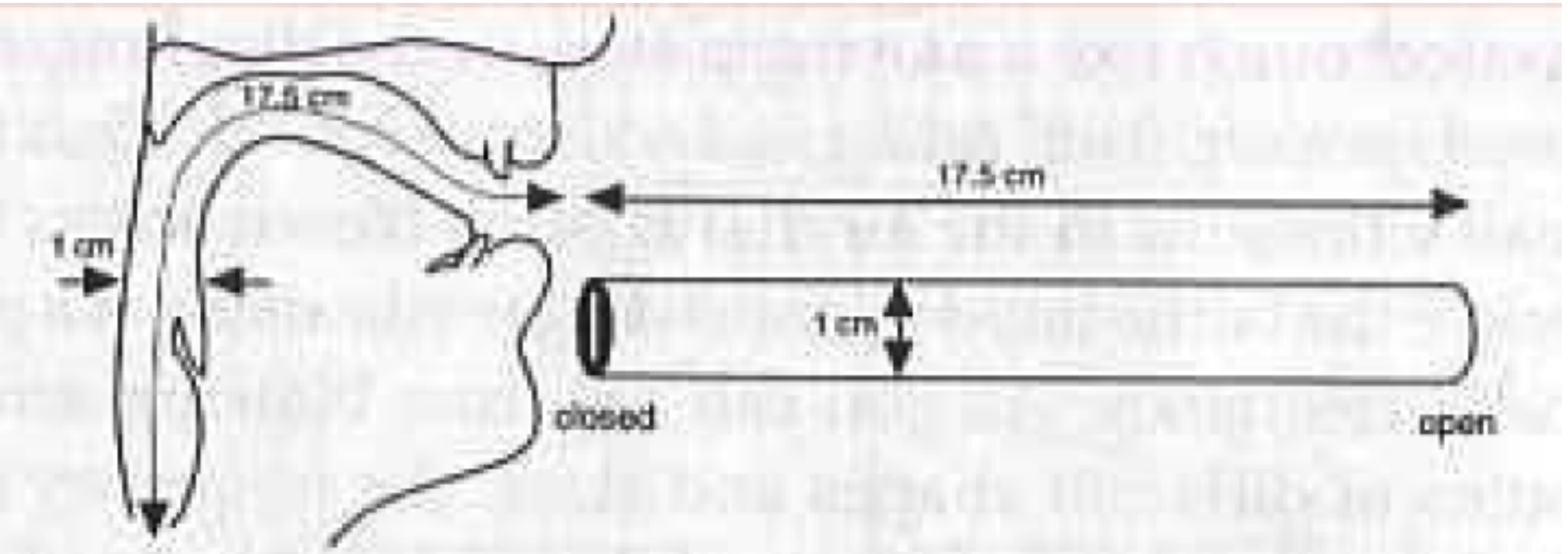


3.
4.
5.
6.
7.
8.
9.

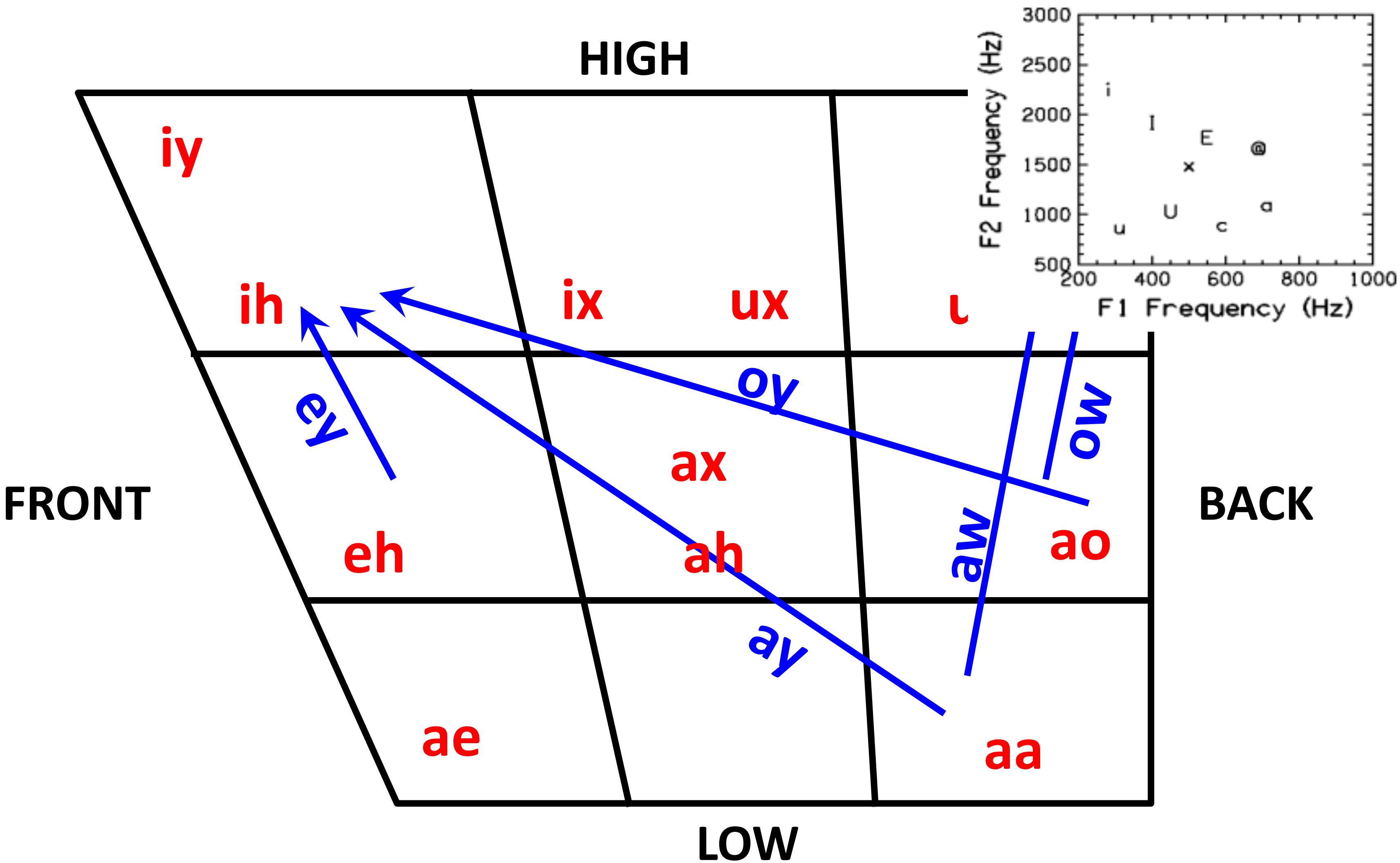
From Ladefoged “A Course in Phonetics”

Deriving Schwa

- Reminder of basic facts about sound waves
  - f = c/λ
  - c = speed of sound (approx 35,000 cm/sec)
  - A sound with λ=10 meters: f = 35 Hz (35,000/1000)
  - A sound with λ=2 centimeters: f = 17,500 Hz (35,000/2)



American English Vowel Space



Figures from Jennifer Venditti, H. T. Bunnell

Dialect Issues

- Speech varies from dialect to dialect (examples are American vs. British English)
  - Syntactic ("I could" vs. "I could do")
  - Lexical ("elevator" vs. "lift")
  - Phonological
  - Phonetic

- Mismatch between training and testing dialects can cause a large increase in error rate