Algorithms for NLP

Acoustic Models

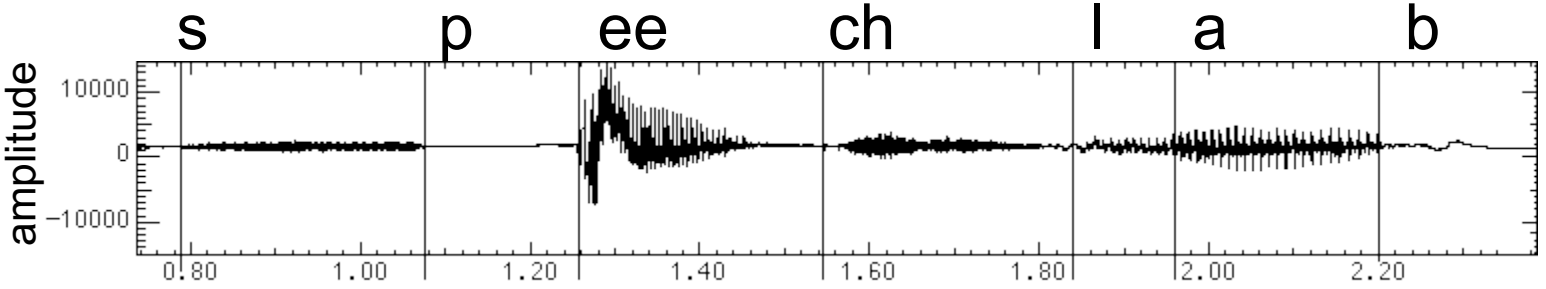
Taylor Berg-Kirkpatrick – CMU

Slides: Dan Klein – UC Berkeley

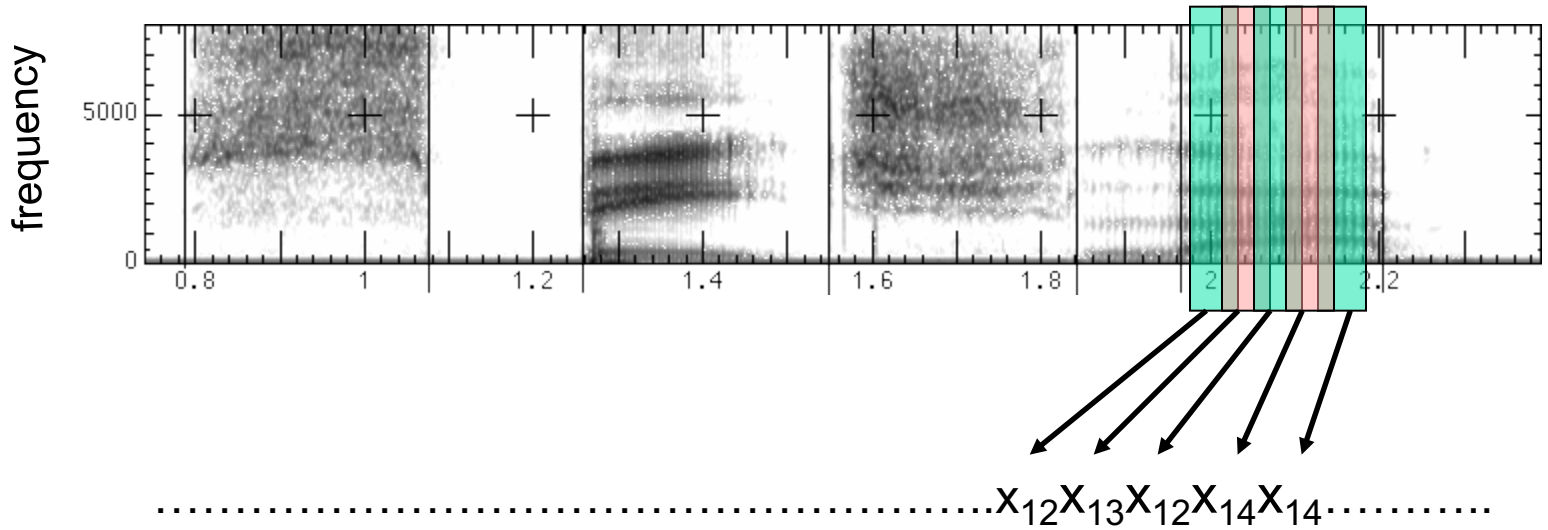
Speech Signals

Speech in a Slide

- Frequency gives pitch; amplitude gives volume

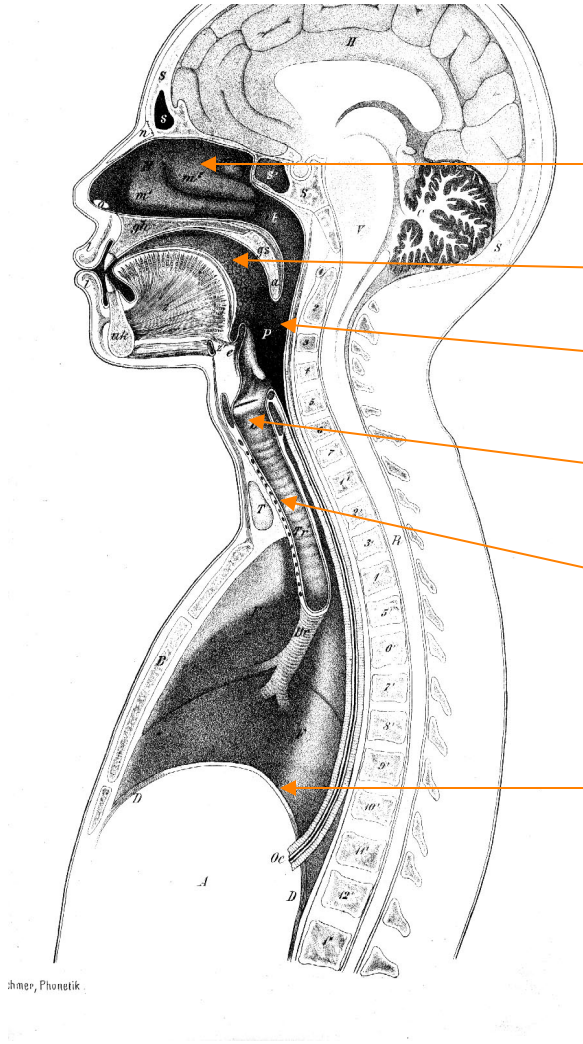


- Frequencies at each time slice processed into observation vectors



Articulation

Articulatory System



Nasal cavity

Oral cavity

Pharynx

Vocal folds (in the larynx)

Trachea

Lungs

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	ɸ β		t d		ʈ ɖ	c ɟ	k g	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		ʡ	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ								
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		ɭ̆						

- Standard international phonetic alphabet (IPA) chart of consonants

Place

Places of Articulation

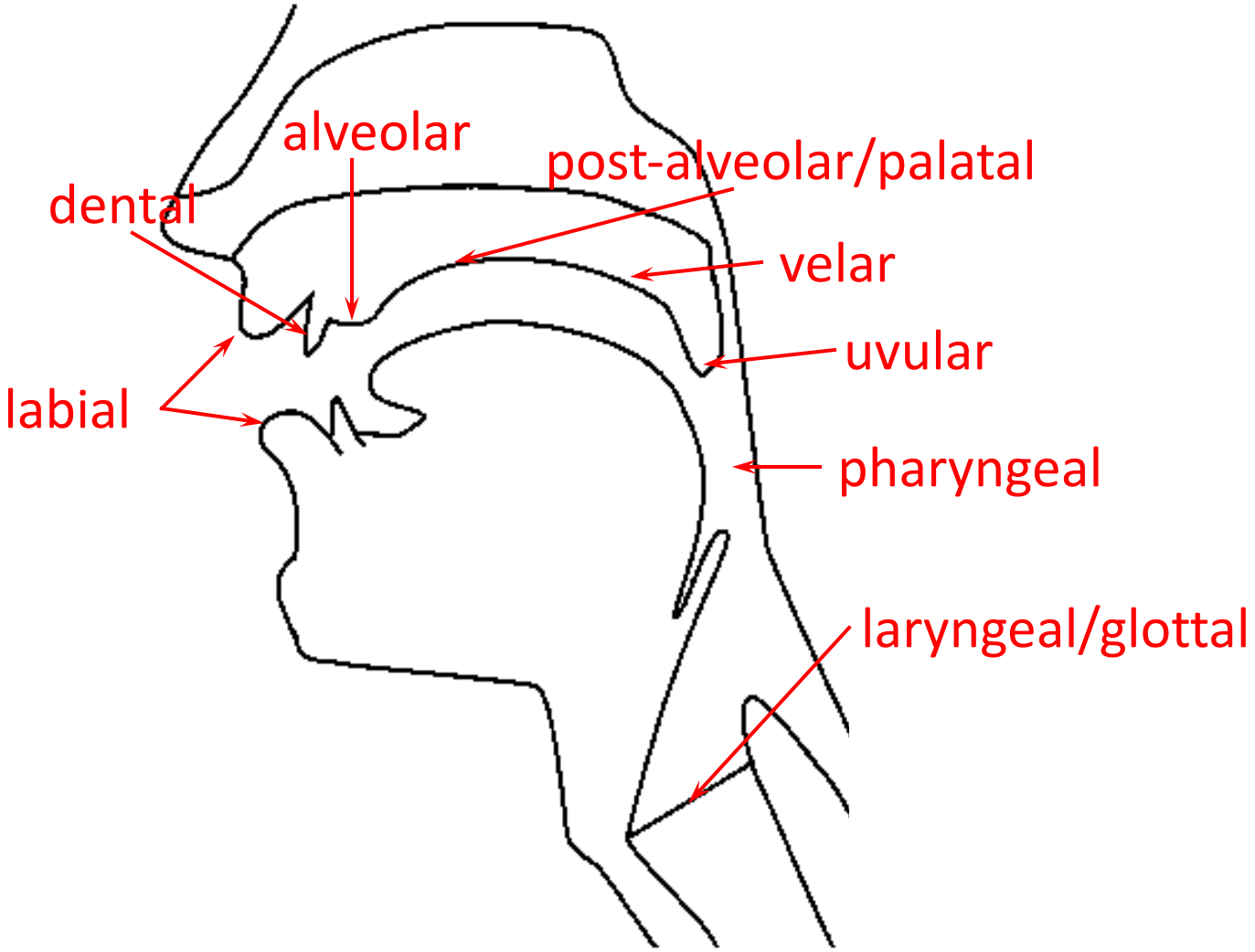
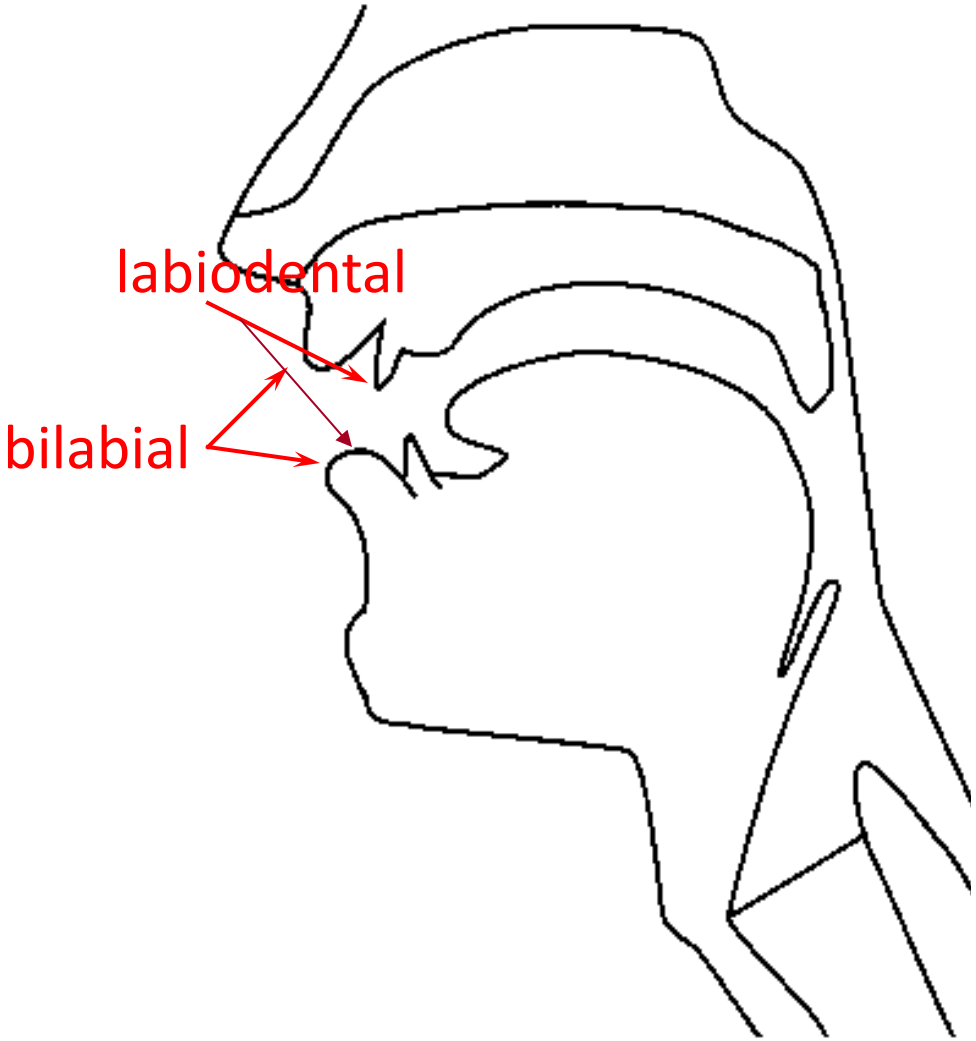


Figure thanks to Jennifer Venditti

Labial place



Bilabial:

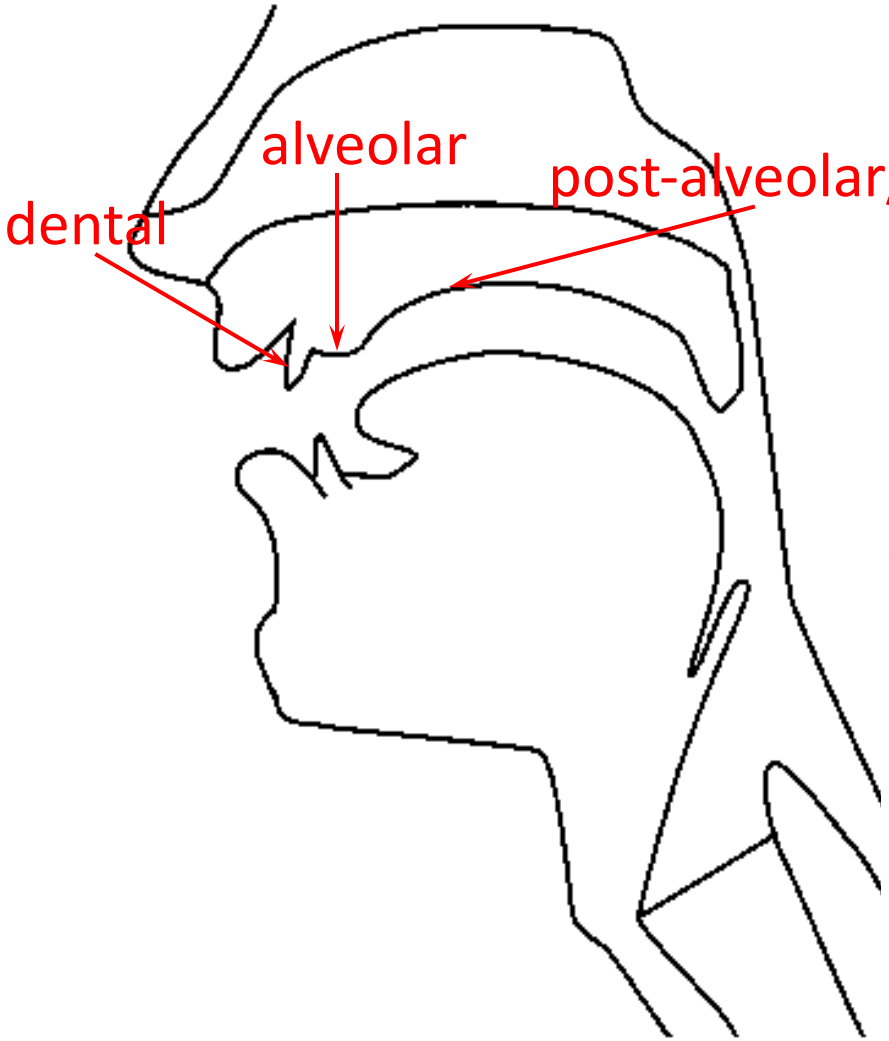
p, b, m

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

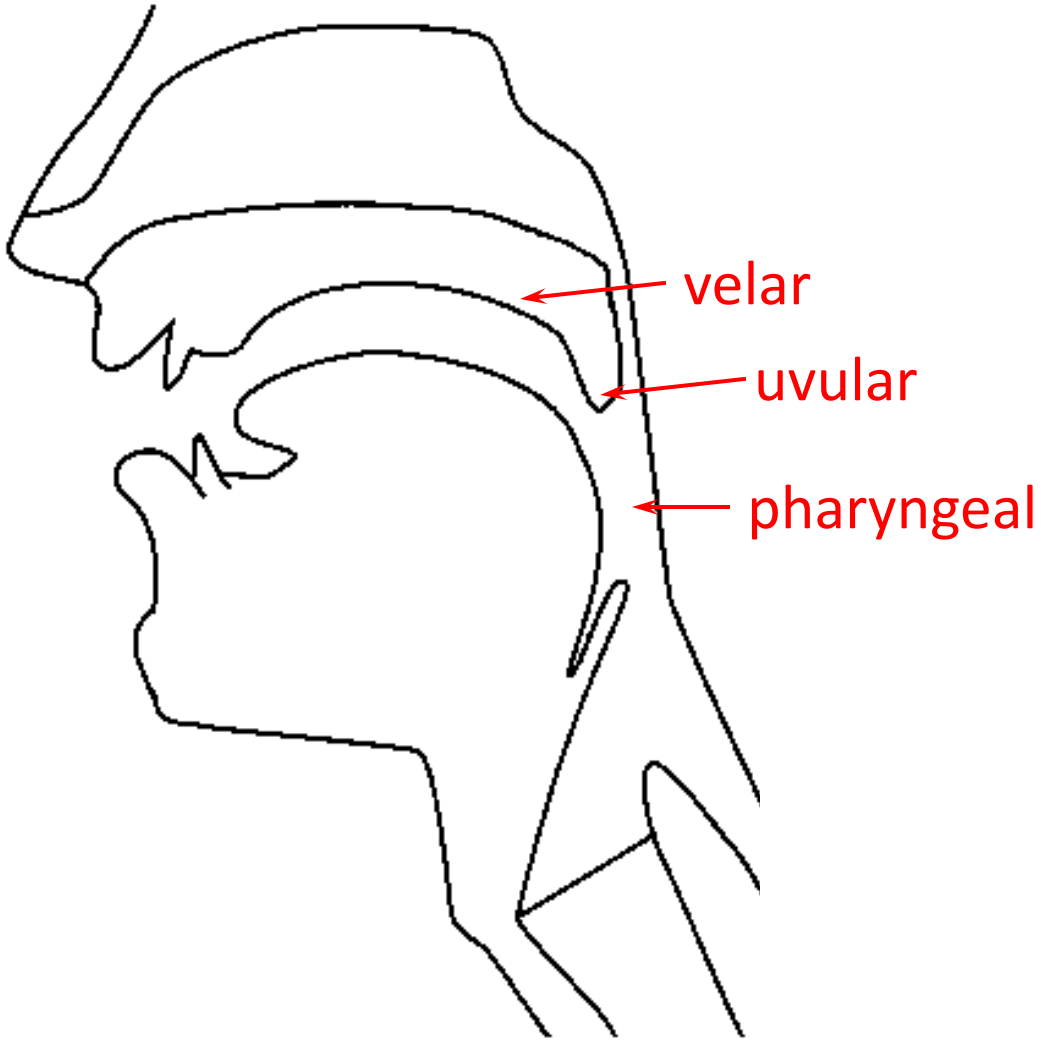


Figure thanks to Jennifer Venditti

Space of Phonemes

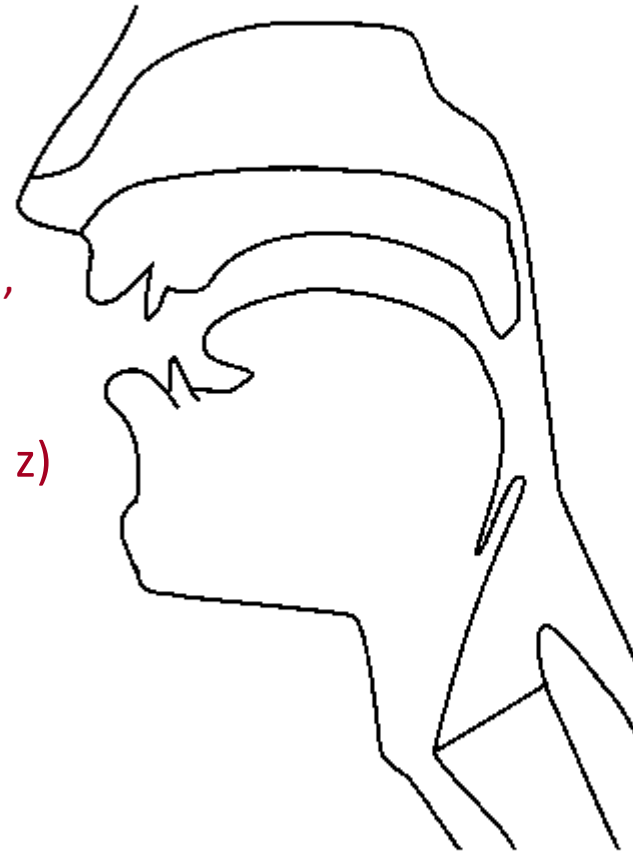
	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ɸ β		t d		ʈ ɖ	c ɟ	k g	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant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

- 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)



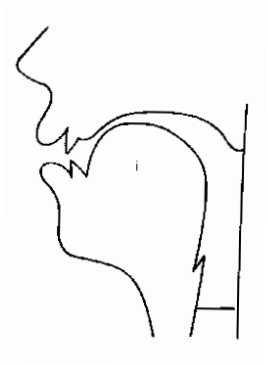
Space of Phonemes

	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	ɸ β		t d		ʈ ɖ	c ɟ	k g	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		ʡ	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ								
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		ɭ̆						

- Standard international phonetic alphabet (IPA) chart of consonants

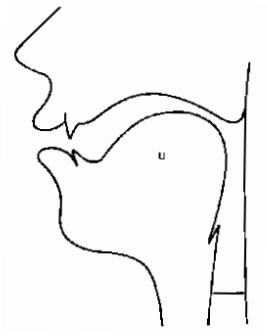
Vowels

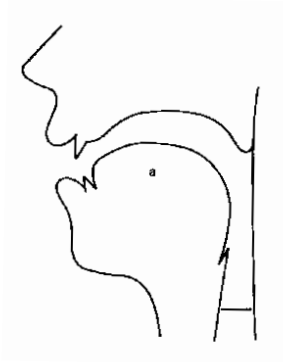
Vowel Space



	Front	Near front	Central	Near back	Back
Close	i	y	ɨ	ʉ	ɯ u
Near close		ɪ • ʏ			• 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

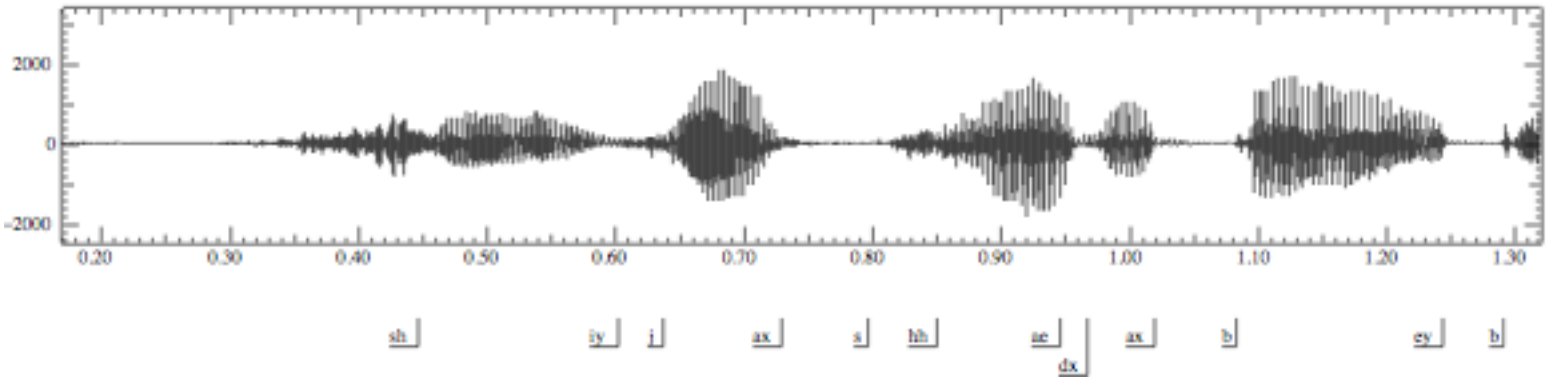
Vowels at right & left of bullets are rounded & unrounded.





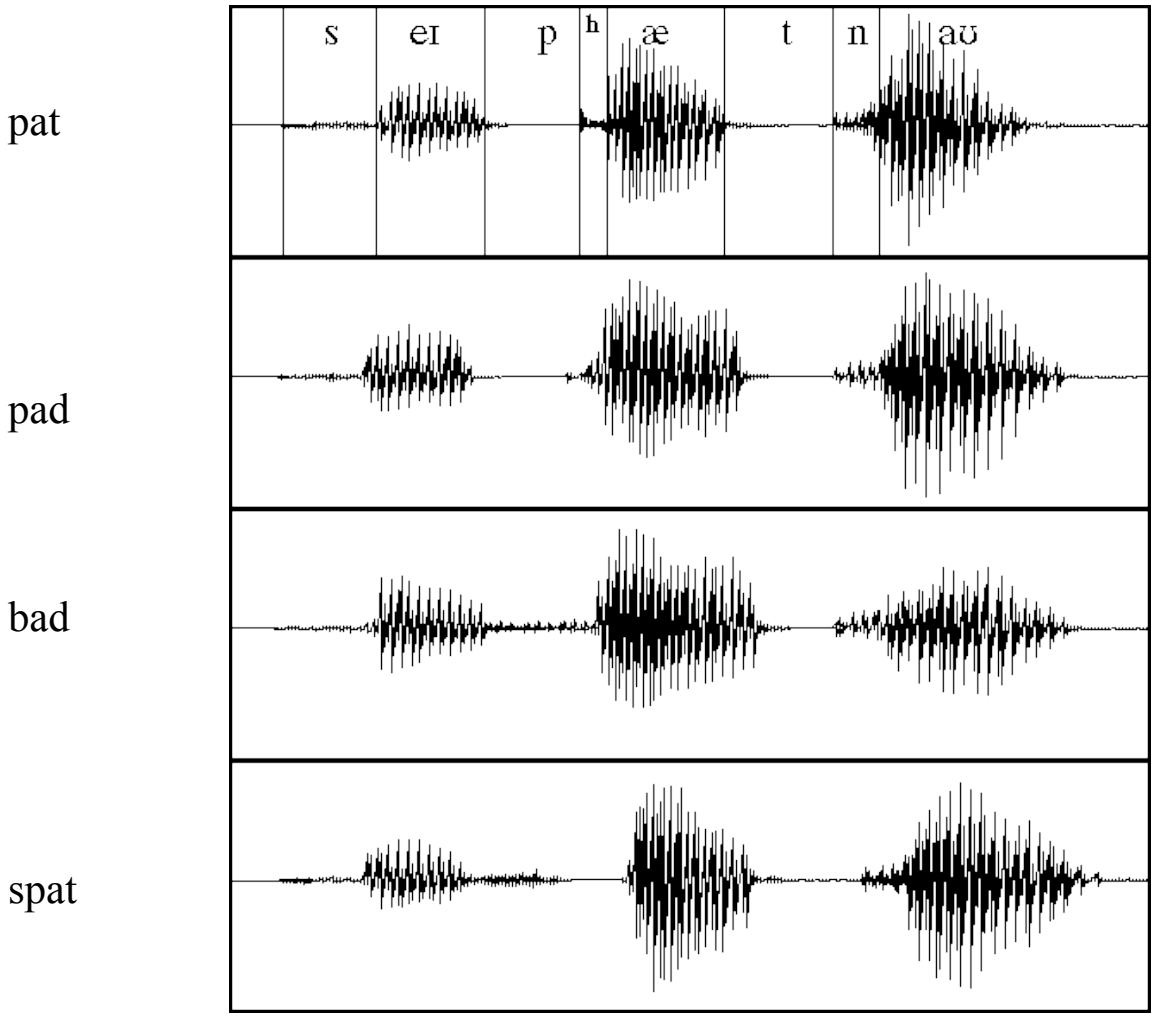
Acoustics

“She just had a baby”



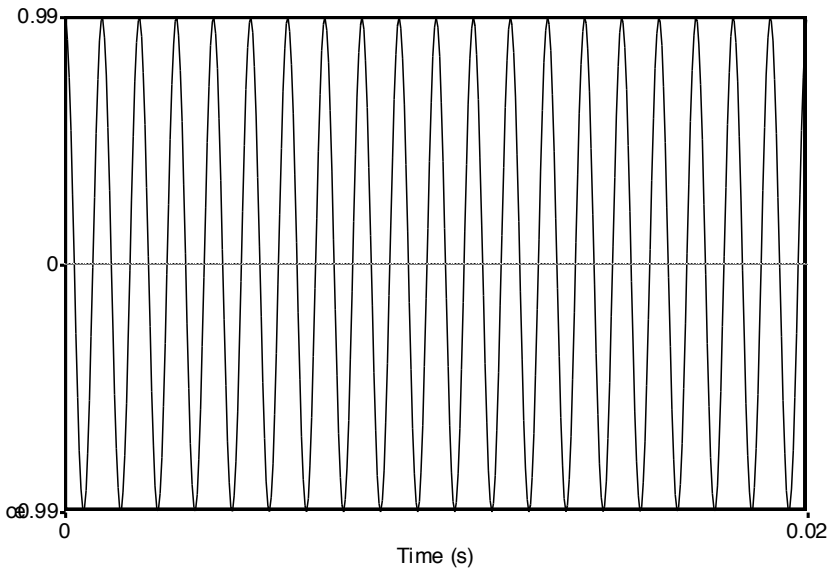
- What can we learn from a wavefile?
  - No gaps between words (!)
  - Vowels are voiced, long, loud
  - Length in time = length in space in waveform picture
  - Voicing: regular peaks in amplitude
  - When stops closed: no peaks, silence
  - Peaks = voicing: .46 to .58 (vowel [iy], from second .65 to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



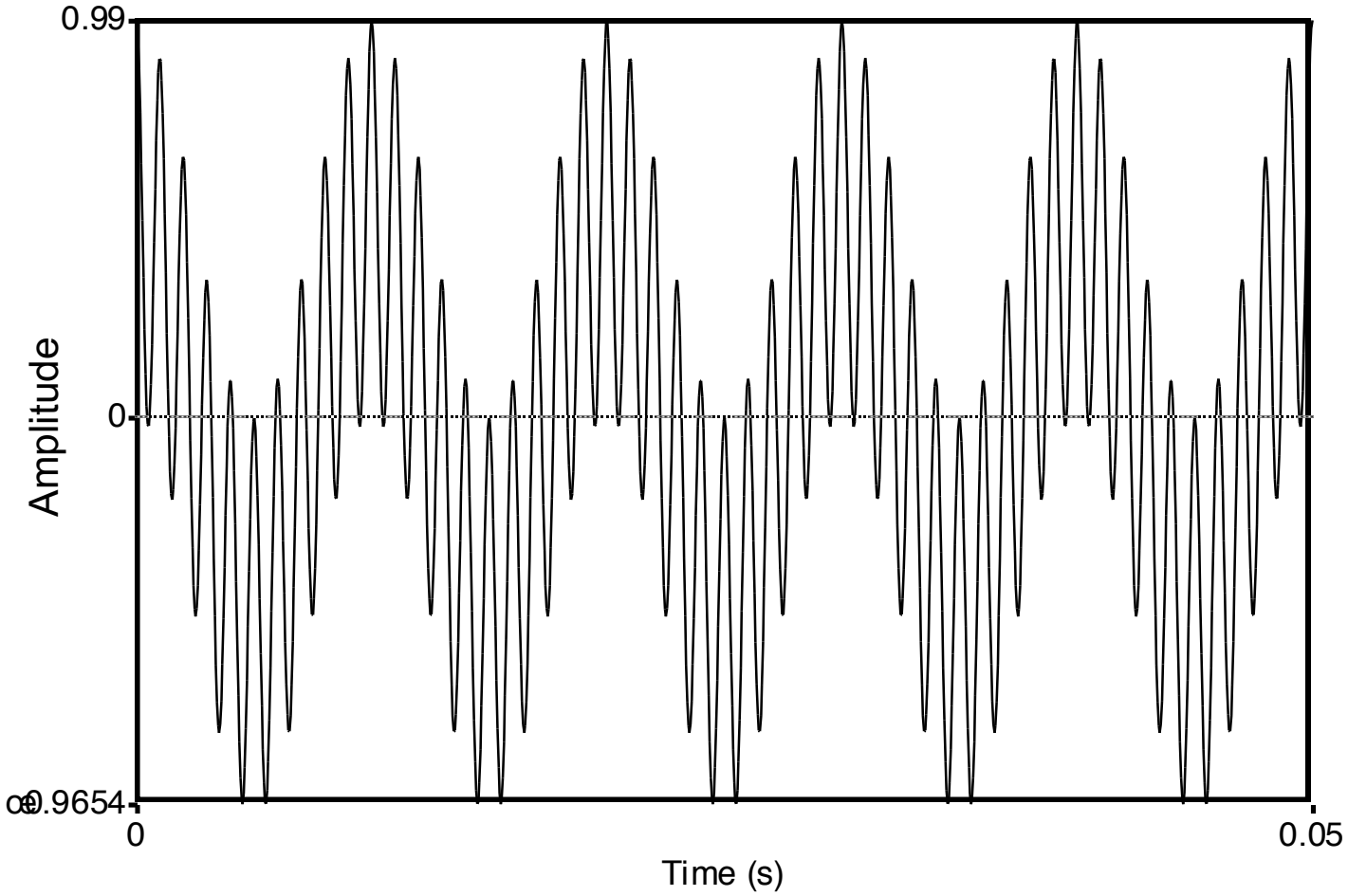
Example from Ladefoged

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

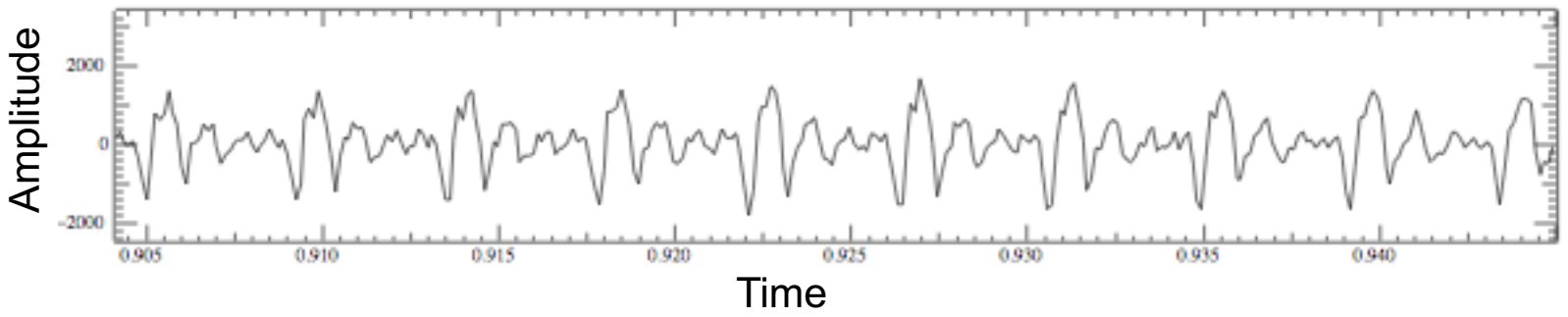
Complex Waves: 100Hz+1000Hz



Spectrum

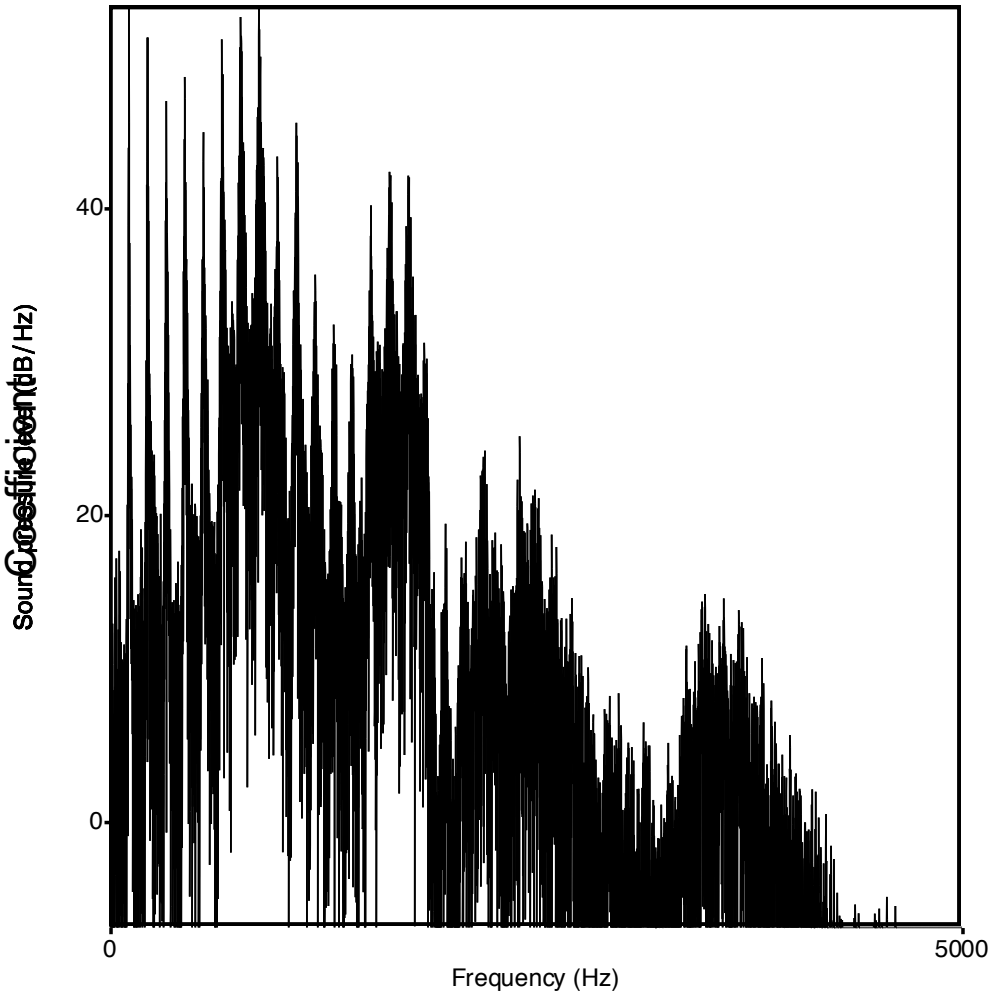
Frequency components (100 and 1000 Hz) on x-axis

Part of [ae] waveform from “had”

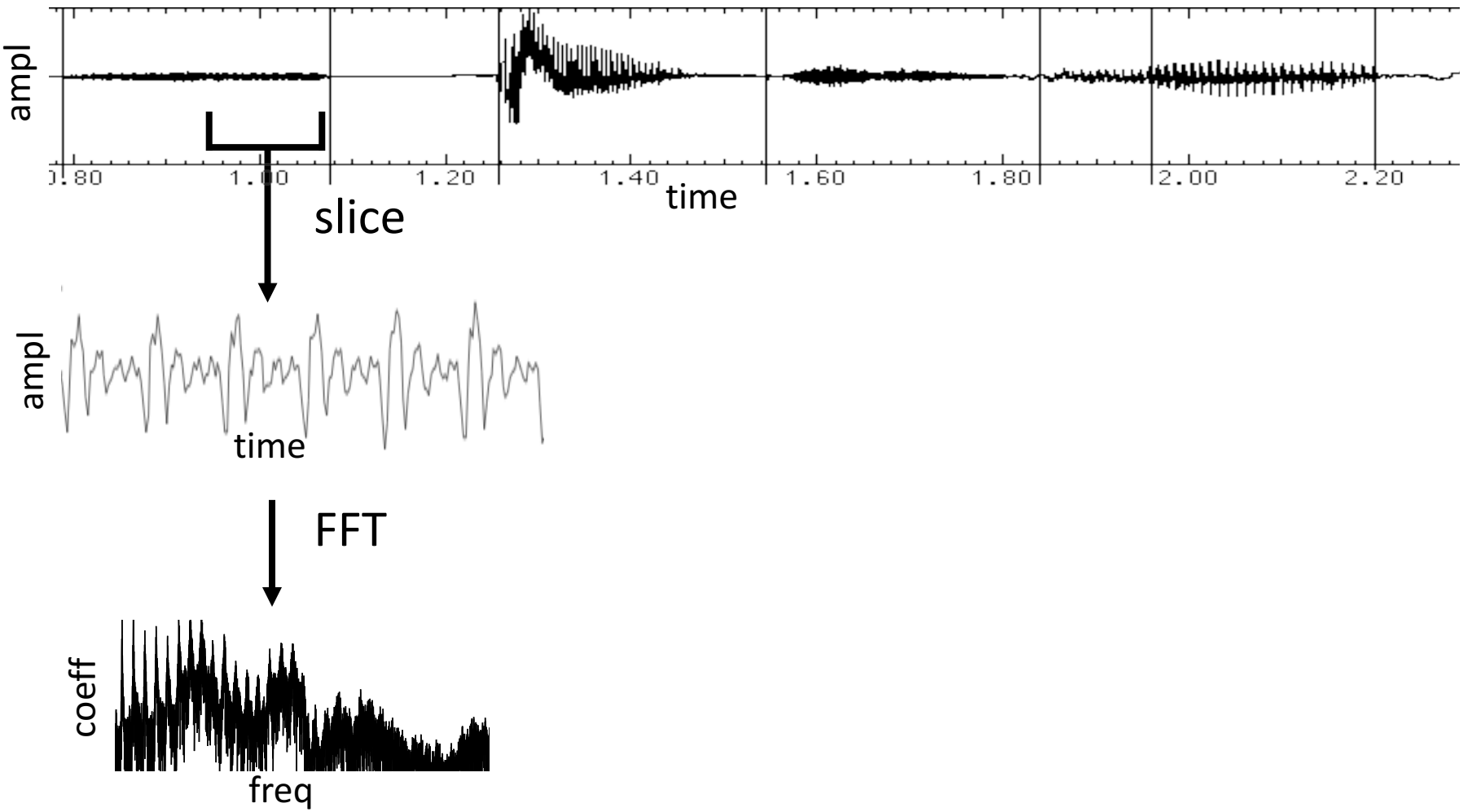


- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves

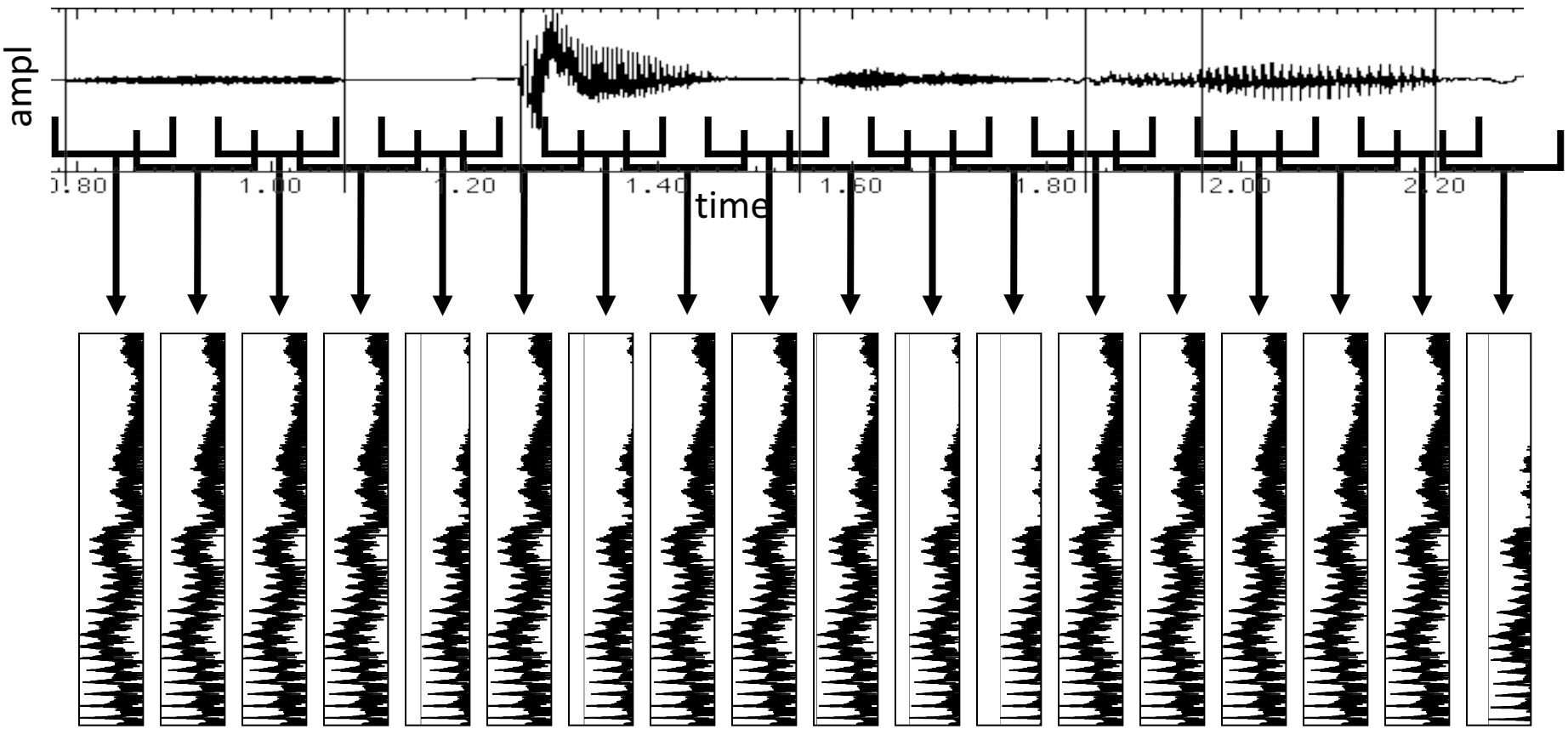
Spectrum of an Actual Speech



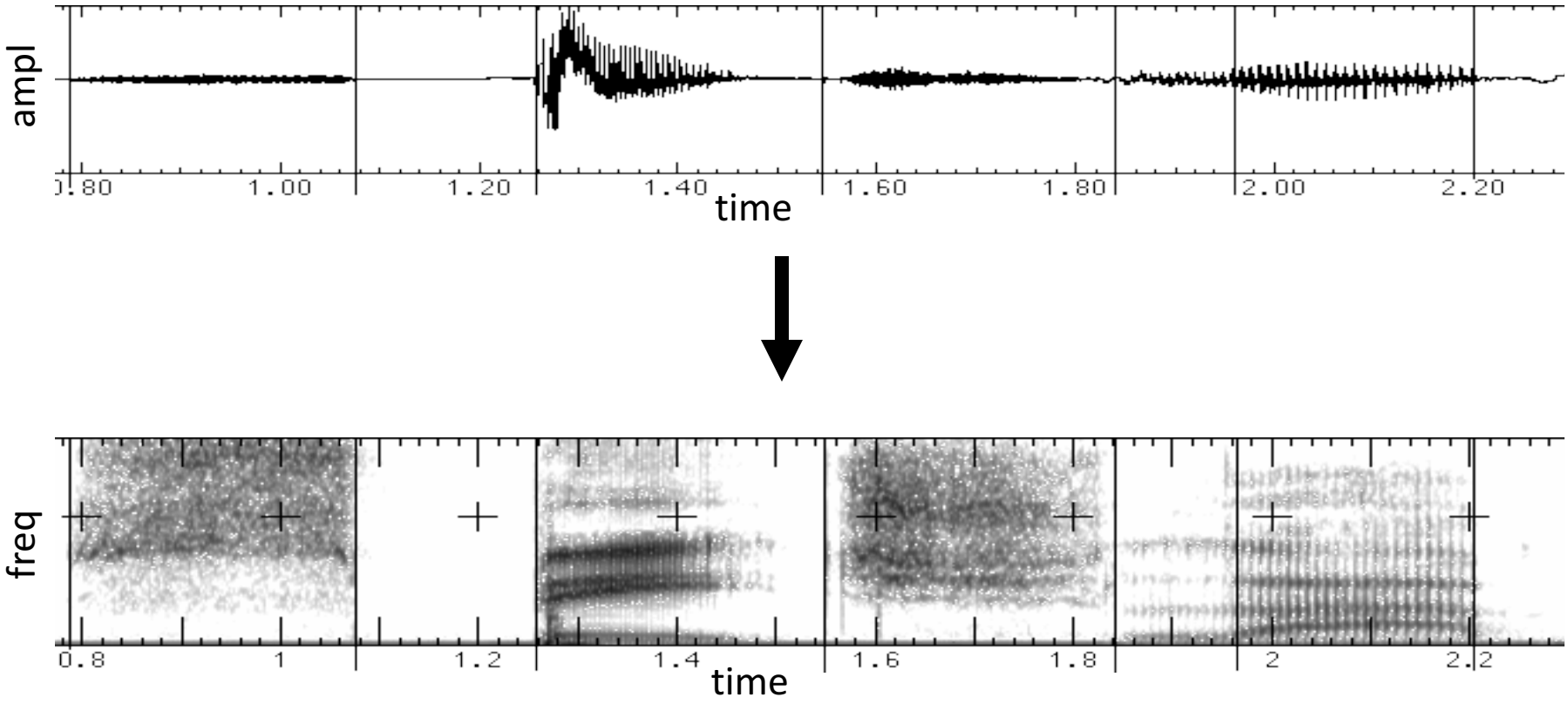
Spectrograms



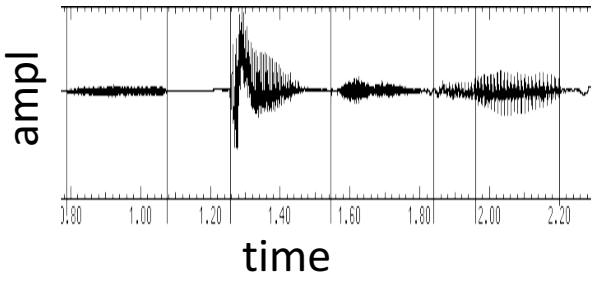
Spectrograms

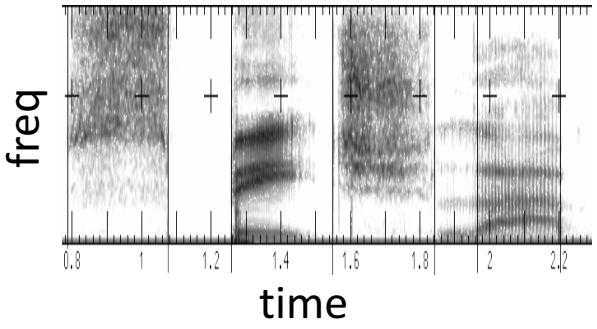


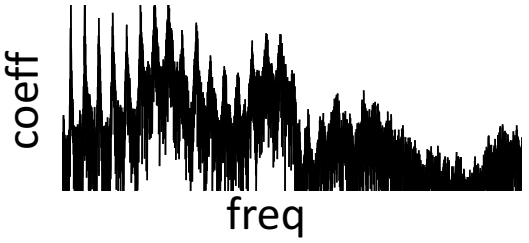
Spectrograms



Types of Graphs

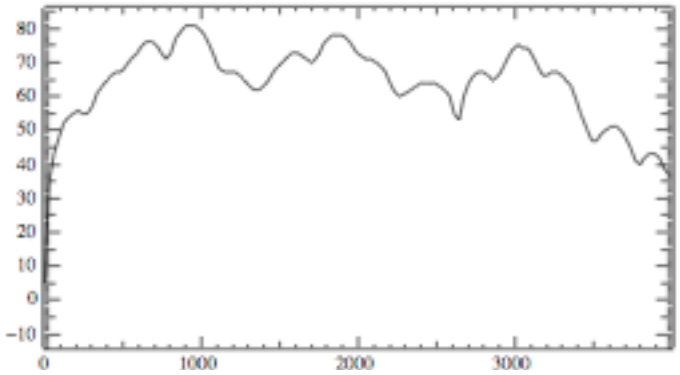






Back to Spectra

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

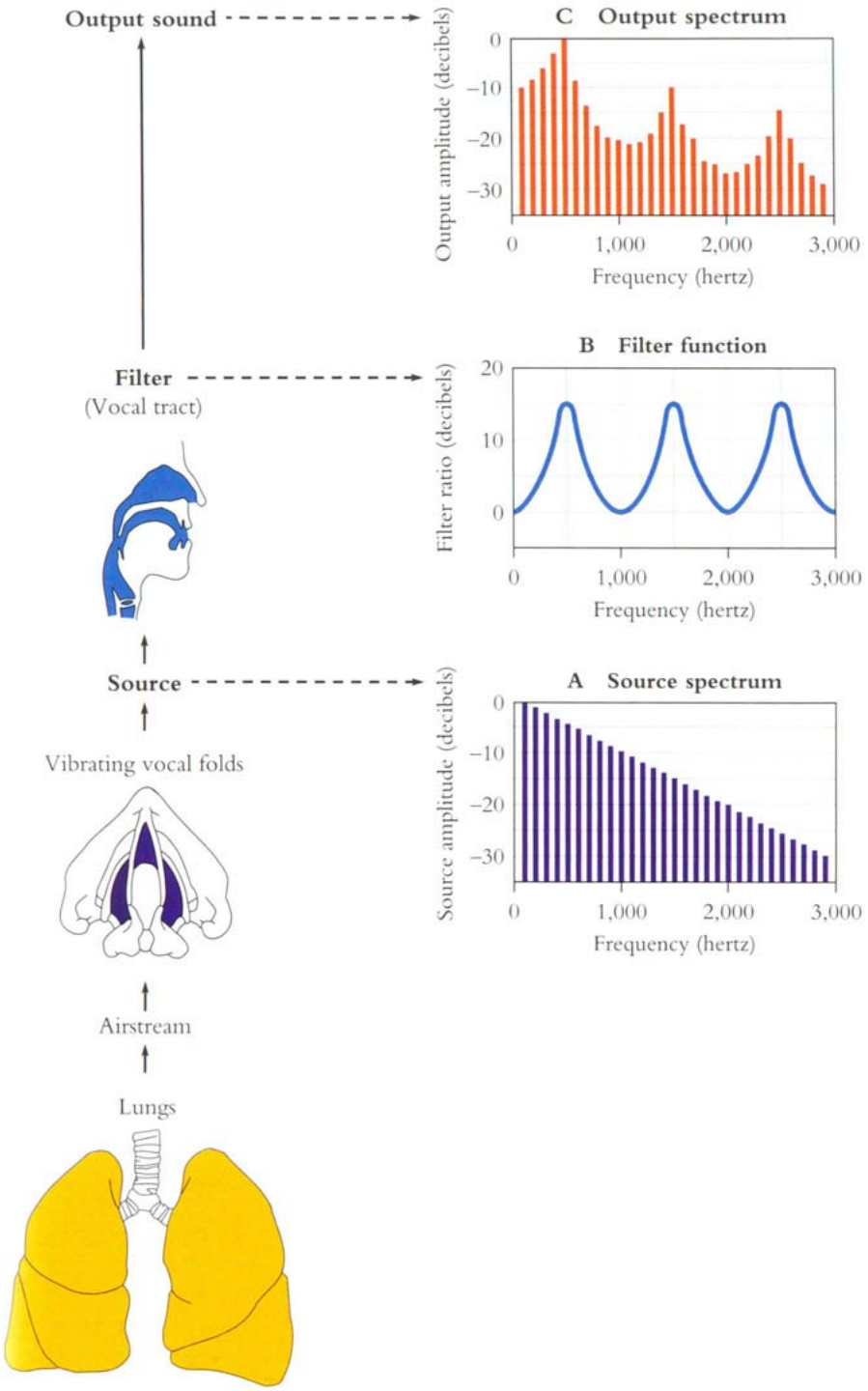


- x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

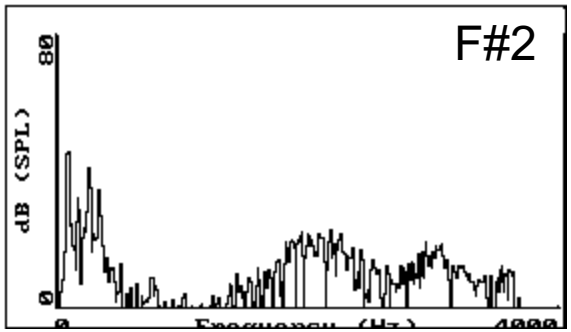
Source / Filter

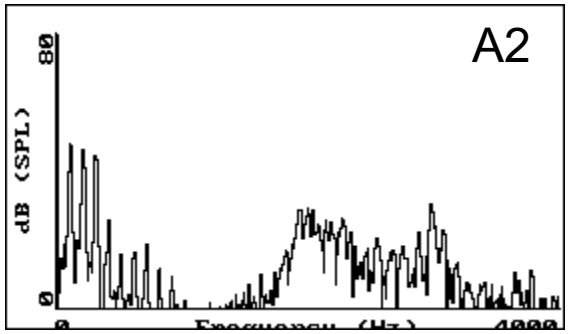
Why these Peaks?

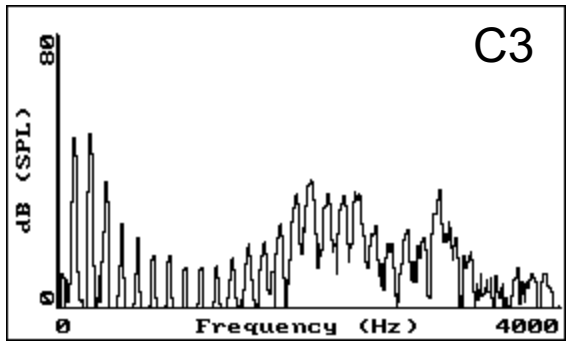
- Articulation process:
  - The vocal cord vibrations create harmonics
  - The mouth is an amplifier
  - Depending on shape of mouth, some harmonics are amplified more than others

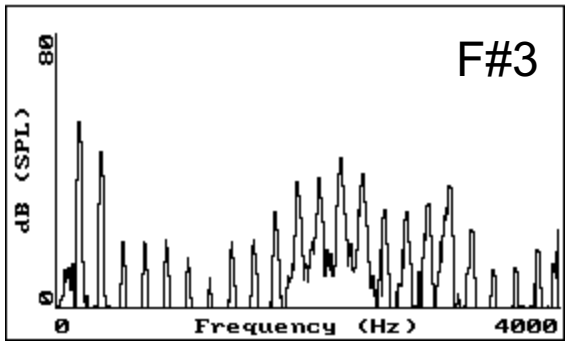


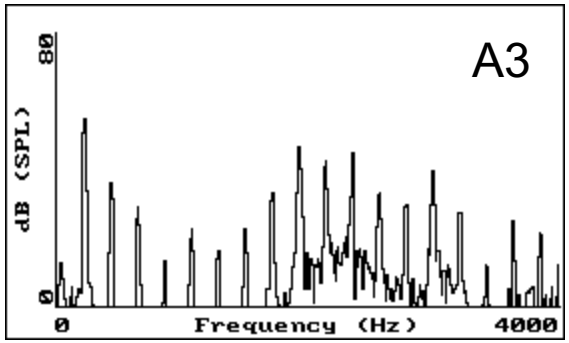
Vowel [i] at increasing pitches

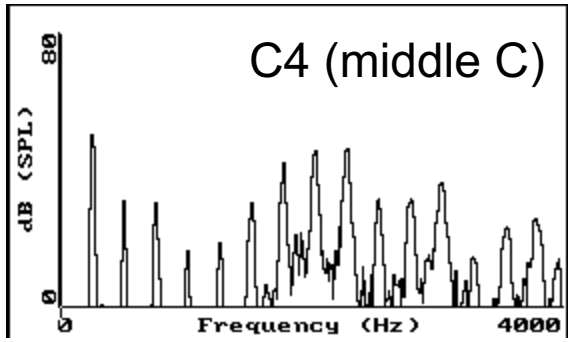


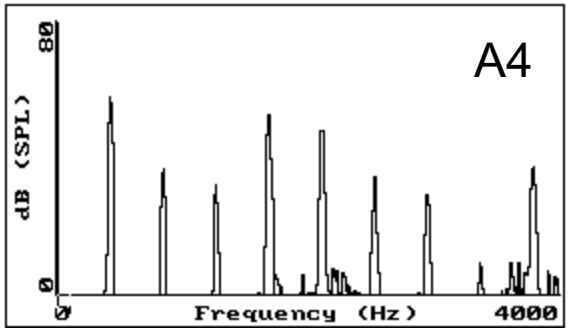








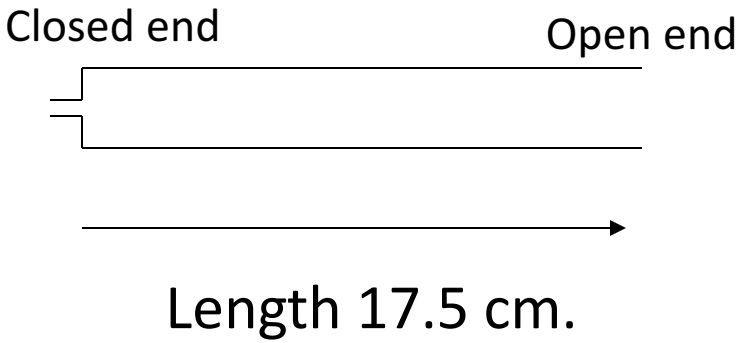


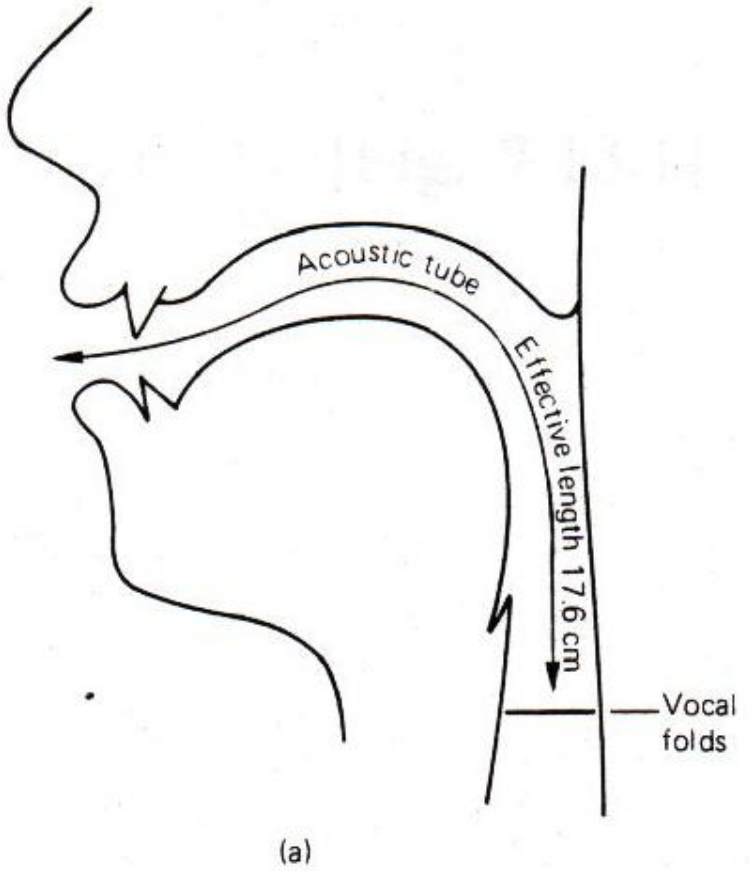


Figures from Ratree Wayland

Resonances of the Vocal Tract

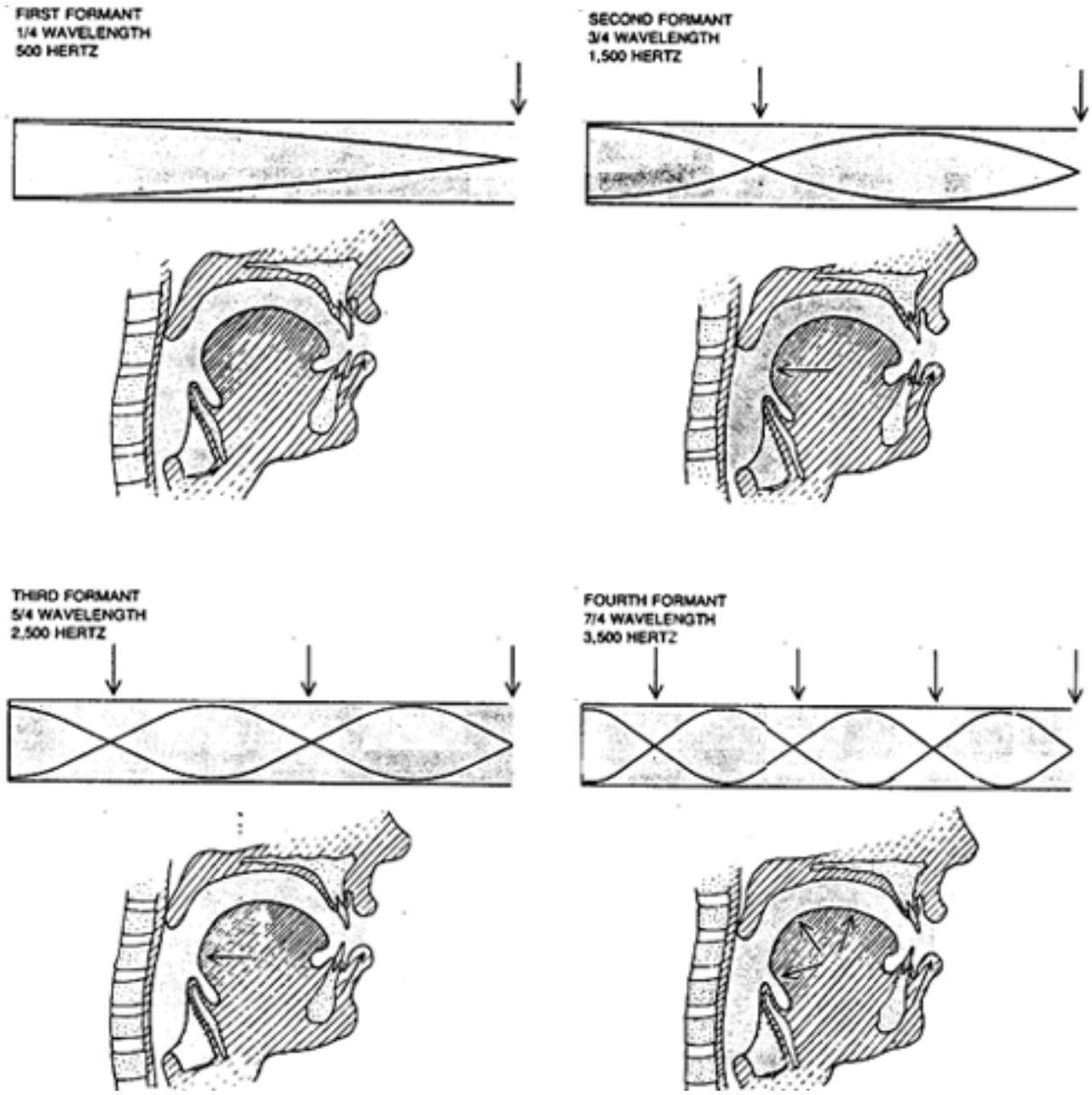
- The human vocal tract as an open tube:





- 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.

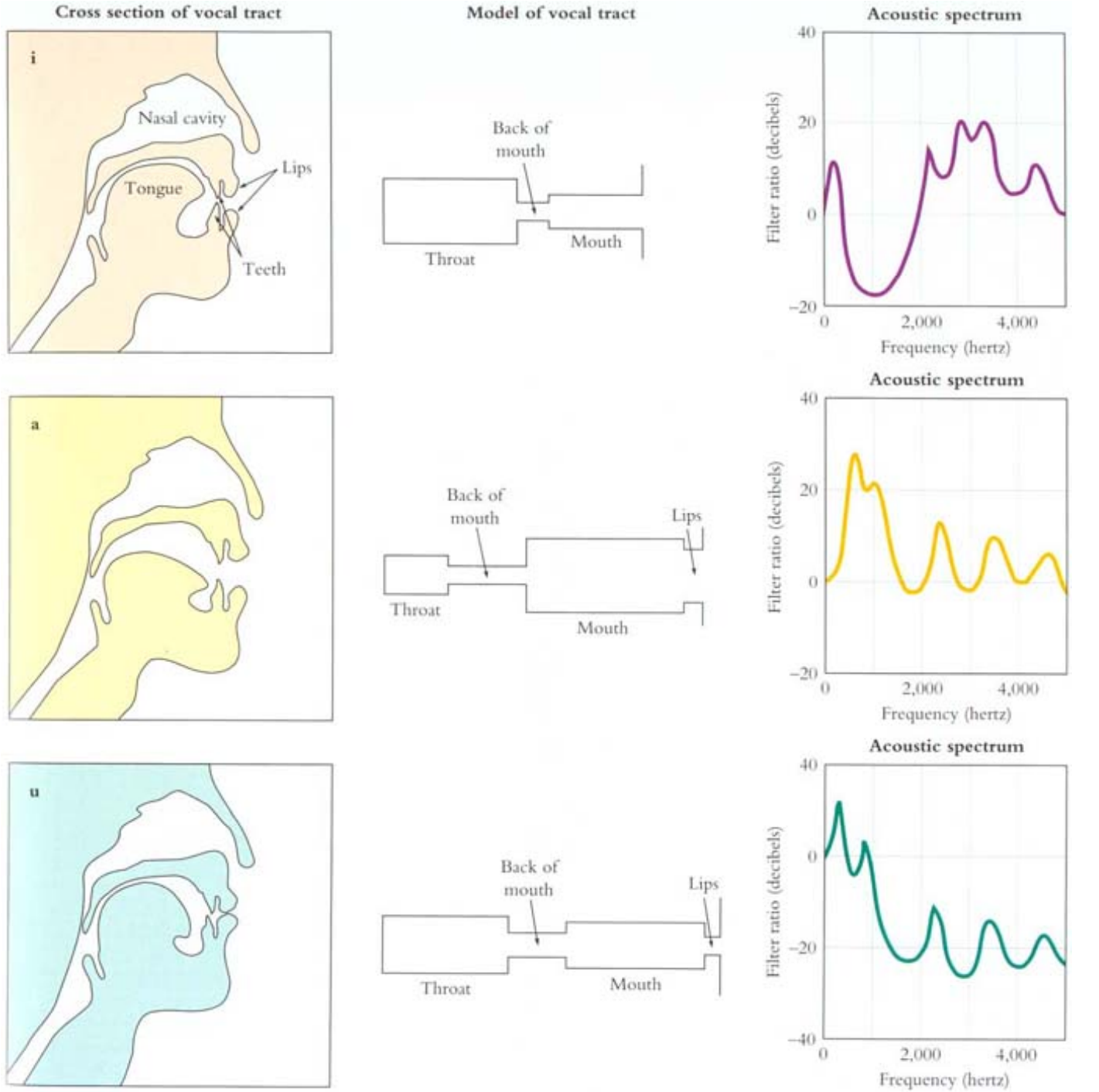
Figure from W. Barry



From Sundberg

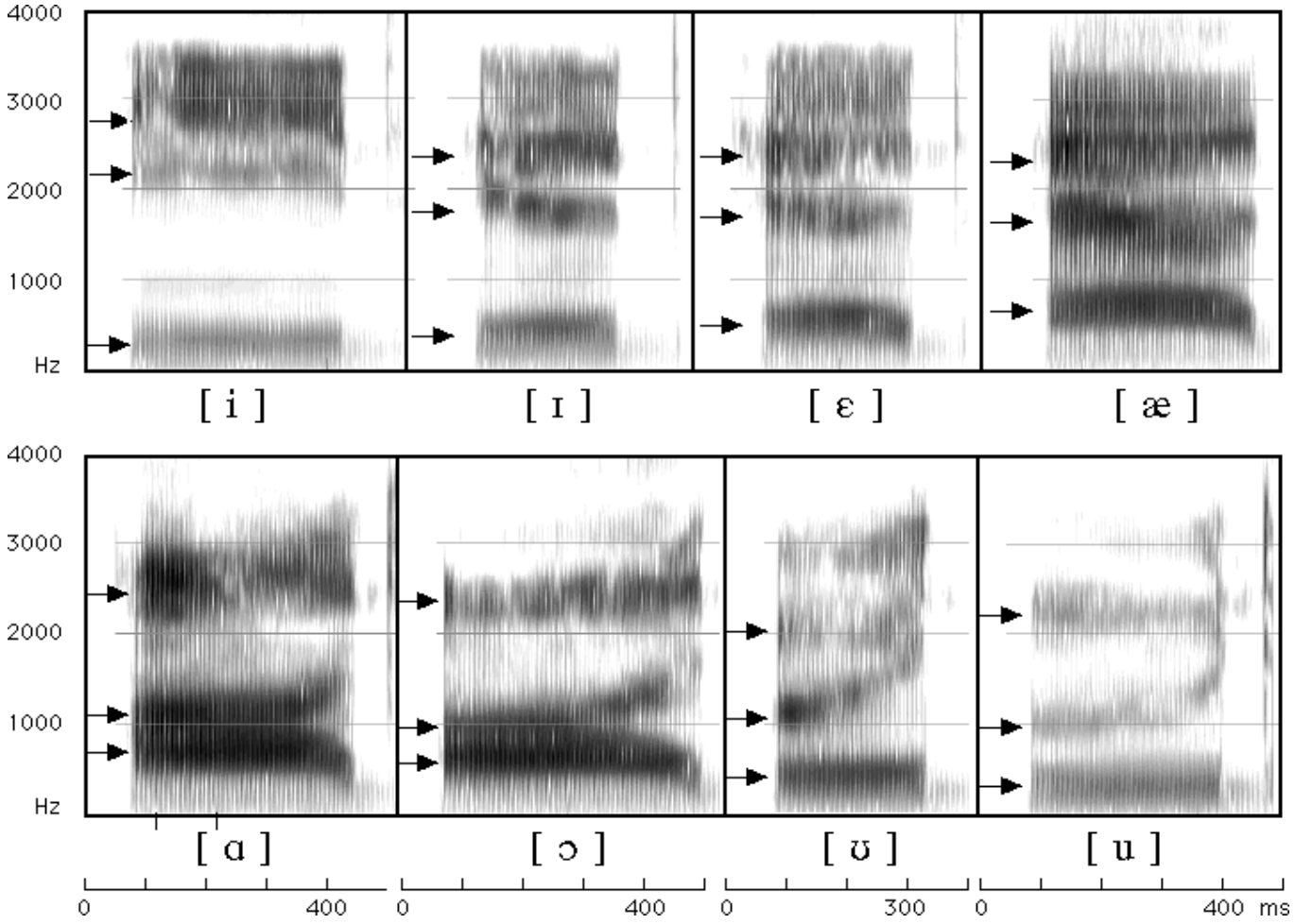
Computing the 3 Formants of Schwa

- 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

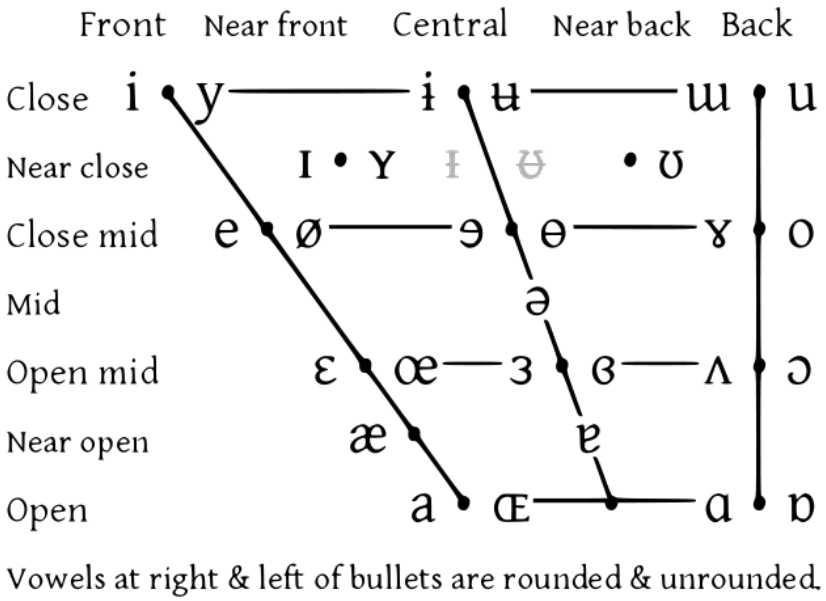


From Mark Liberman

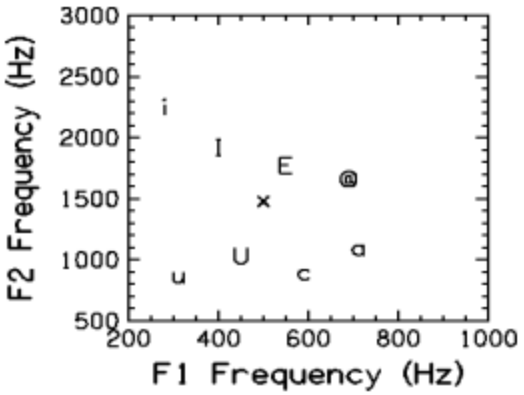
Seeing Formants: the Spectrogram

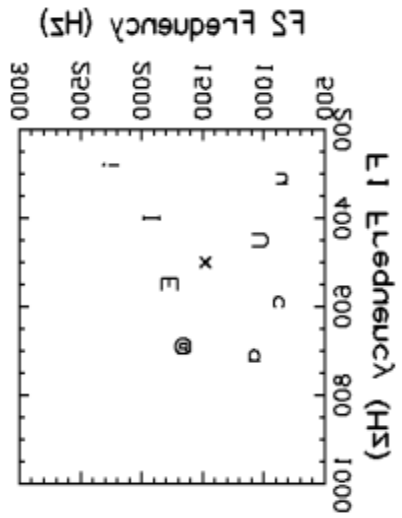


Vowel Space

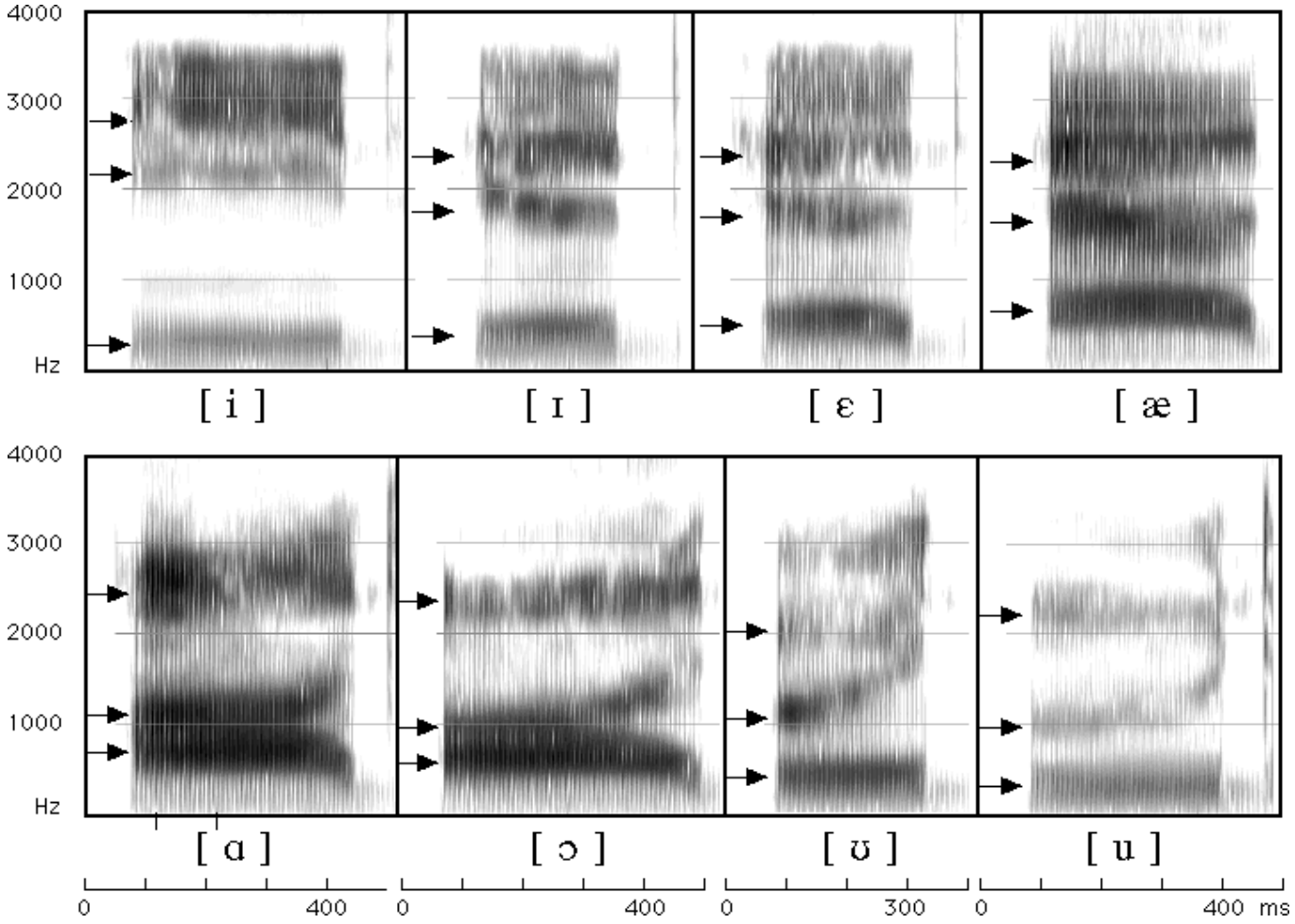


Vowels at right & left of bullets are rounded & unrounded.

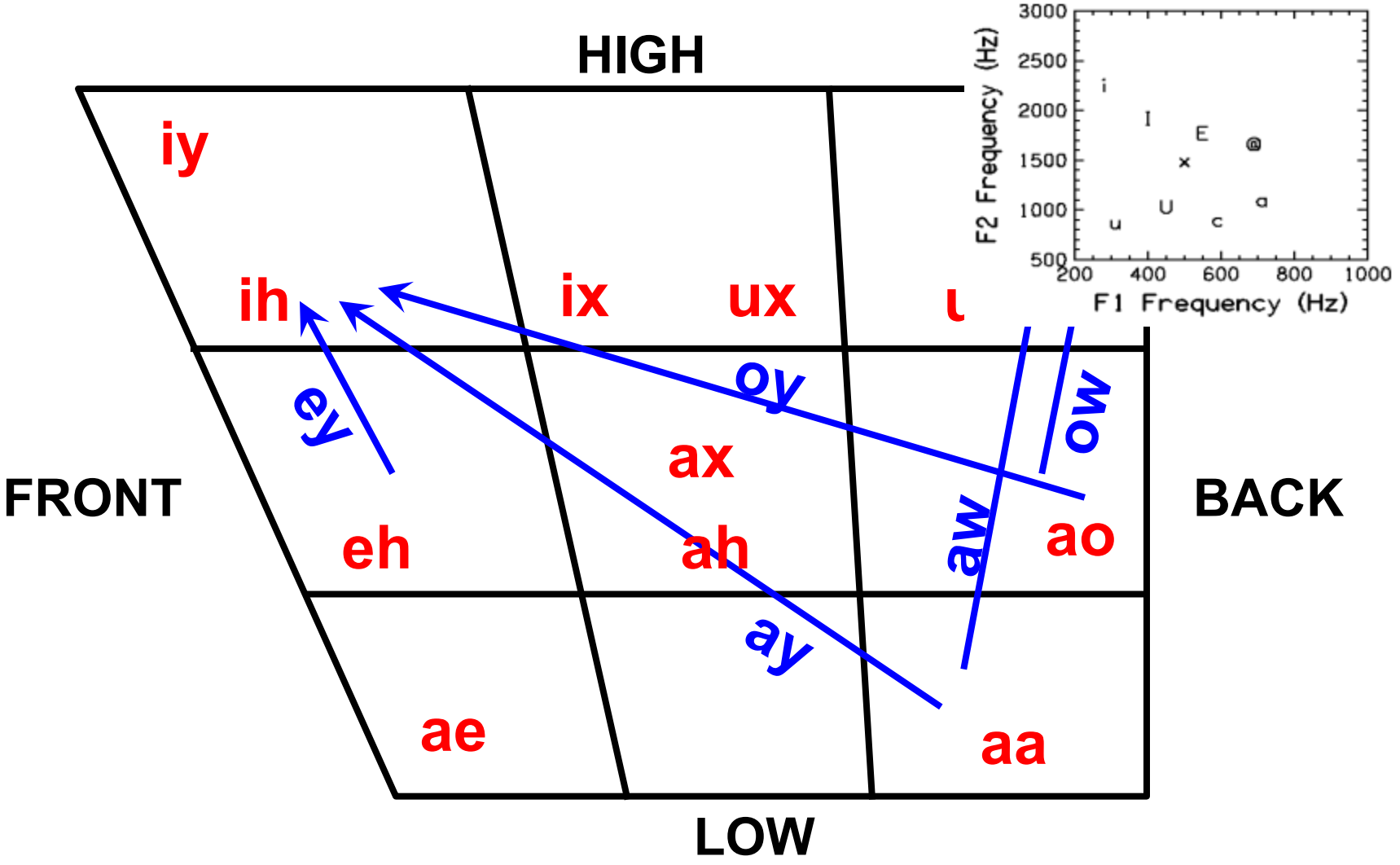




Seeing Formants: the Spectrogram



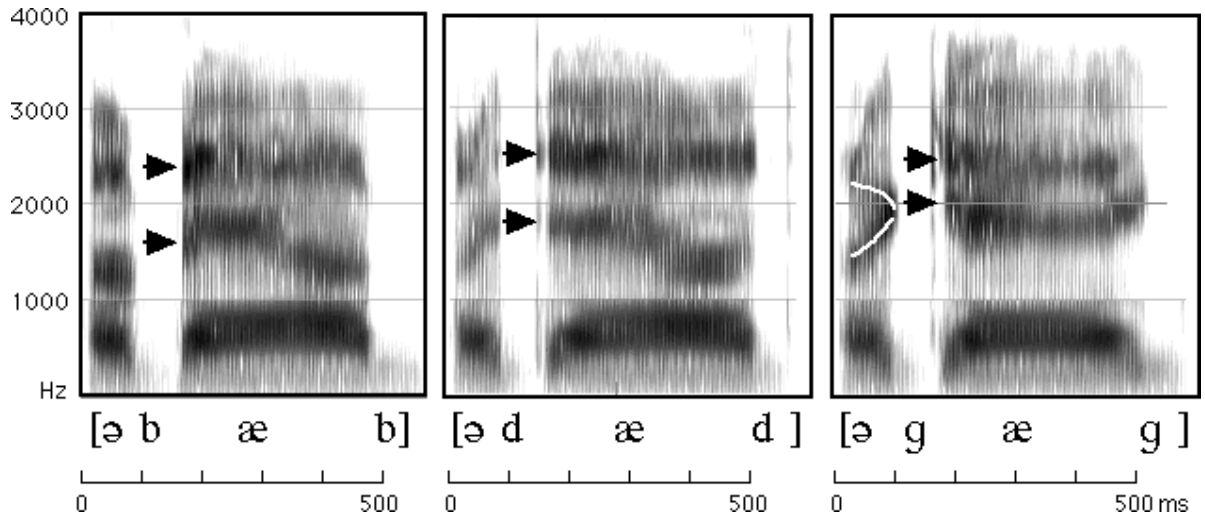
American English Vowel Space



Figures from Jennifer Venditti, H. T. Bunnell

Spectrograms

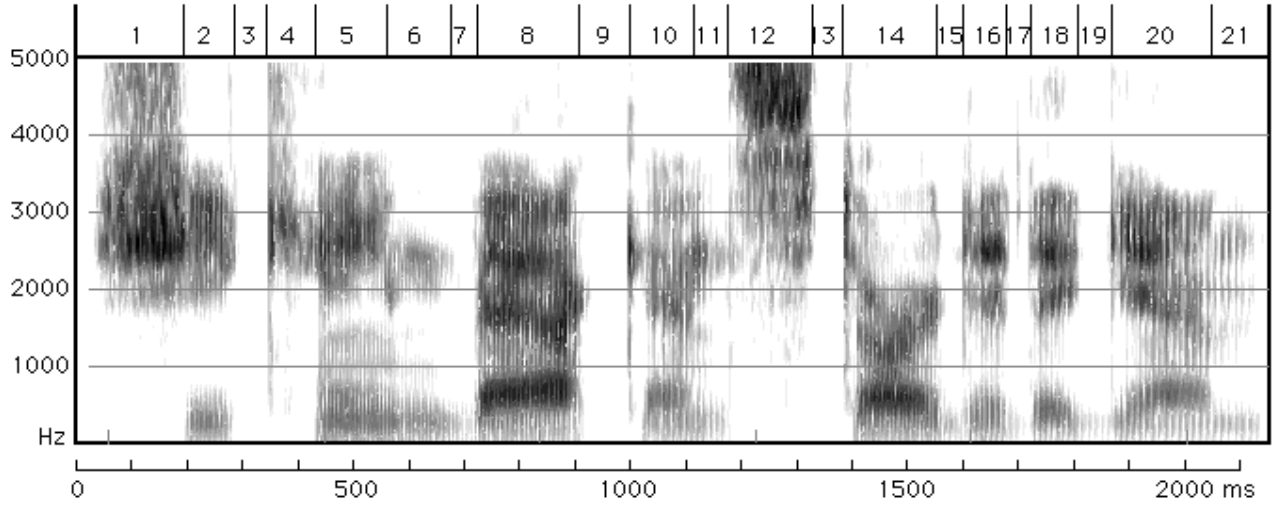
How to Read Spectrograms



- [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”

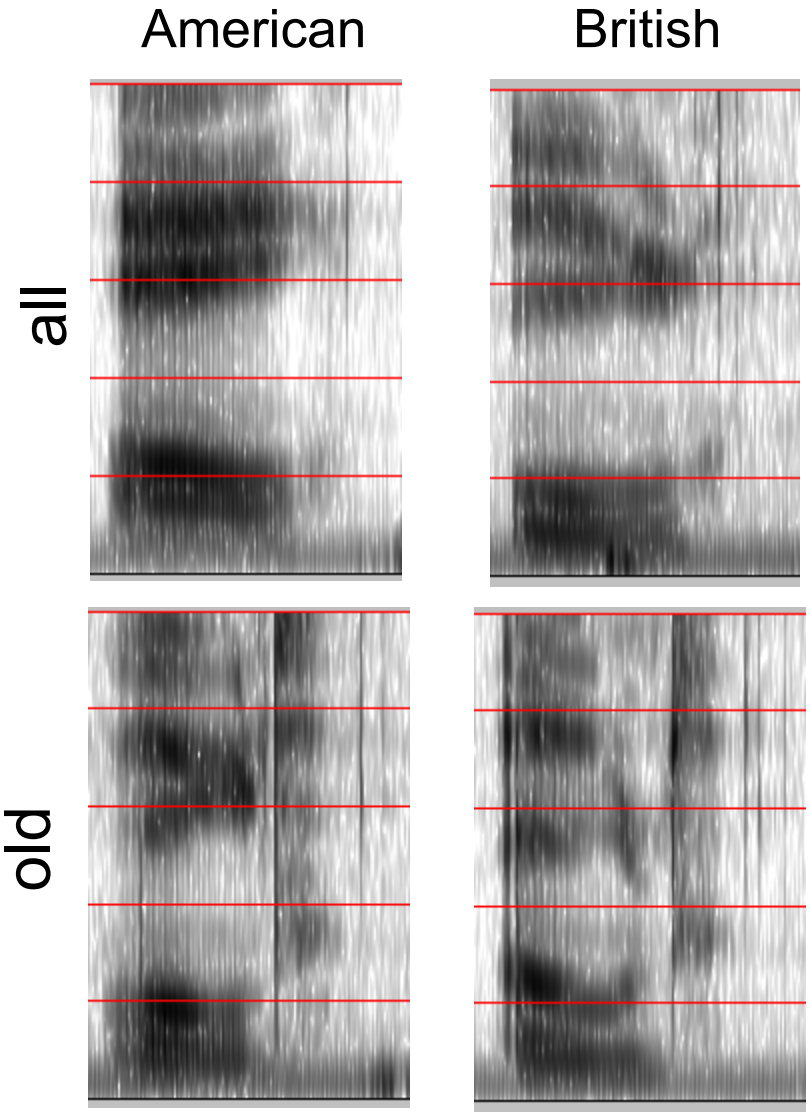


1. lots of high-freq energy
3. closure for k
4. burst of aspiration for k
5. ey vowel; faint 1100 Hz formant is nasalization
6. bilabial nasal
7. short b closure, voicing barely visible.
8. ae; note upward transitions after bilabial stop at beginning
9. note F2 and F3 coming together for "k"

From Ladefoged “A Course in Phonetics”

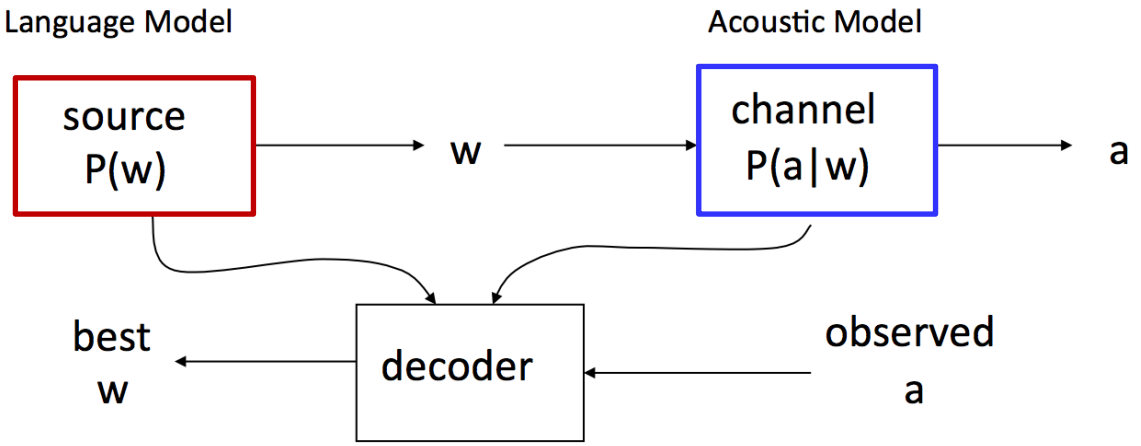
Dialect Issues

- 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



Speech Recognition

The Noisy Channel Model

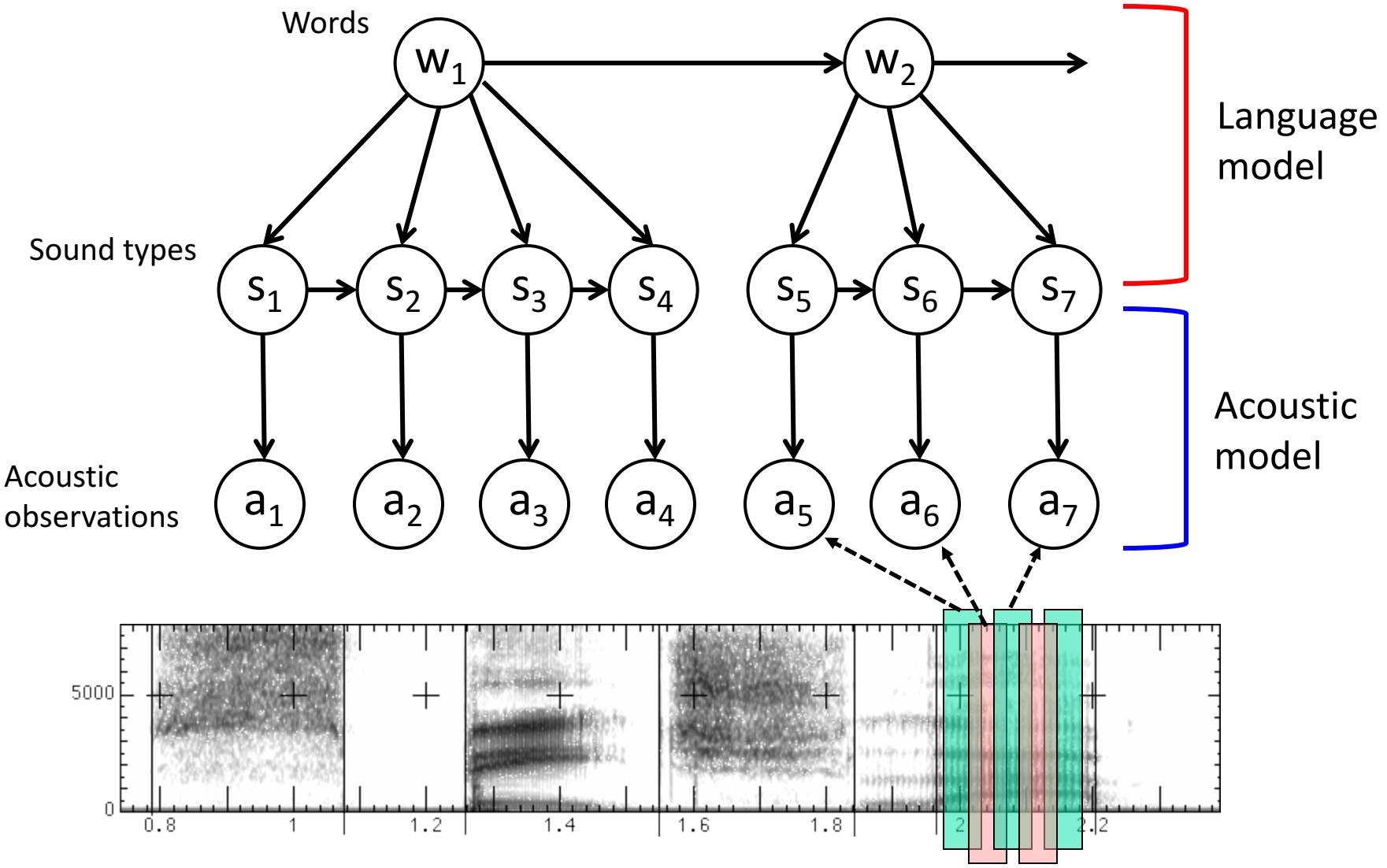


$$w^* = \arg\max_w P(w|a)$$
$$\propto \arg\max_w P(a|w)P(w)$$

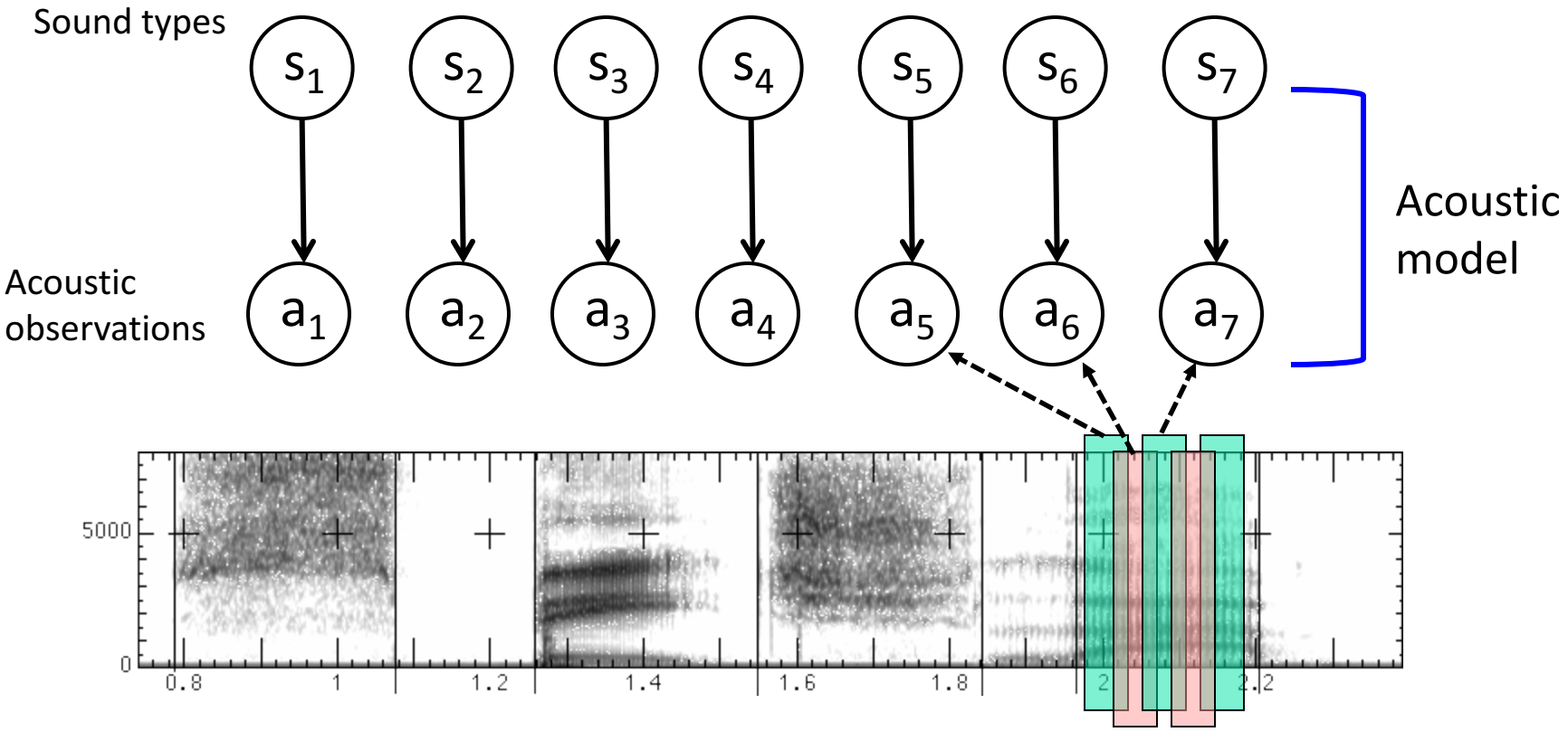
Acoustic model: HMMs over word positions with mixtures of Gaussians as emissions

Language model: Distributions over sequences of words (sentences)

Speech Model

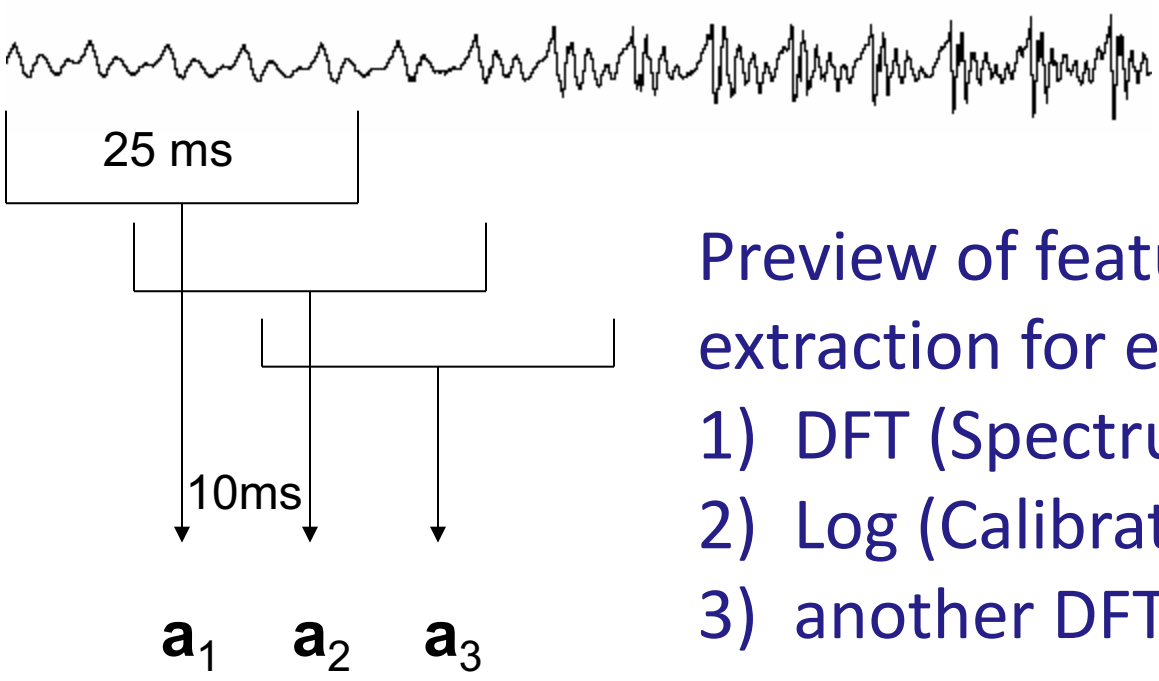


Acoustic Model



Frame Extraction

- A frame (25 ms wide) extracted every 10 ms



Preview of feature extraction for each frame:

1) DFT (Spectrum)
2) Log (Calibrate?)
3) another DFT (!!??)

Figure: Simon Arnfield

Feature Extraction

Digitizing Speech

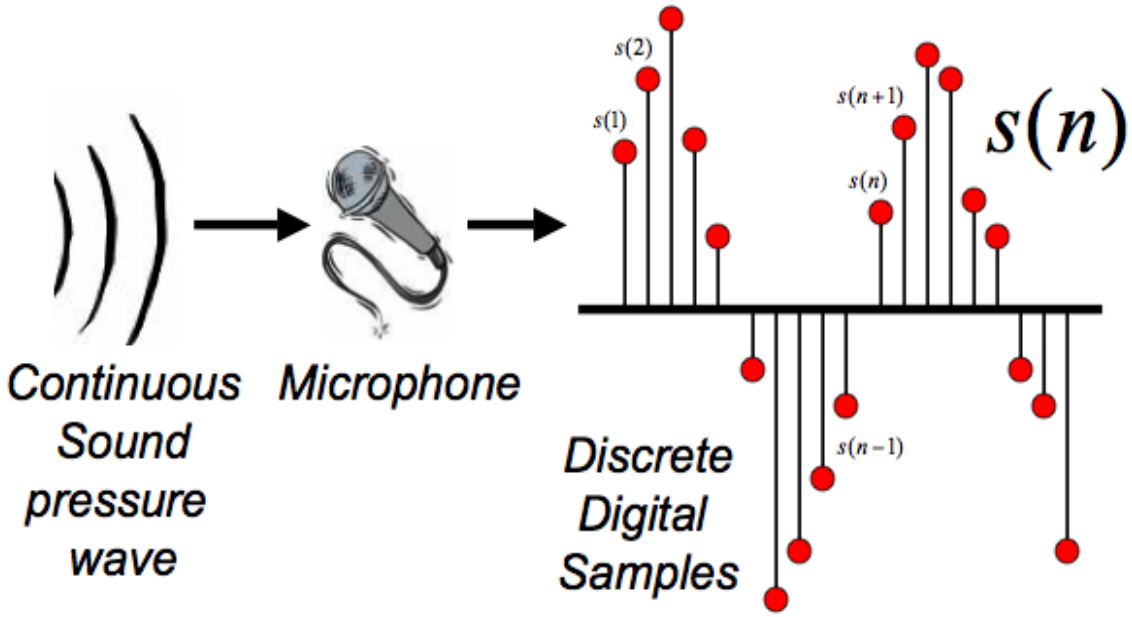
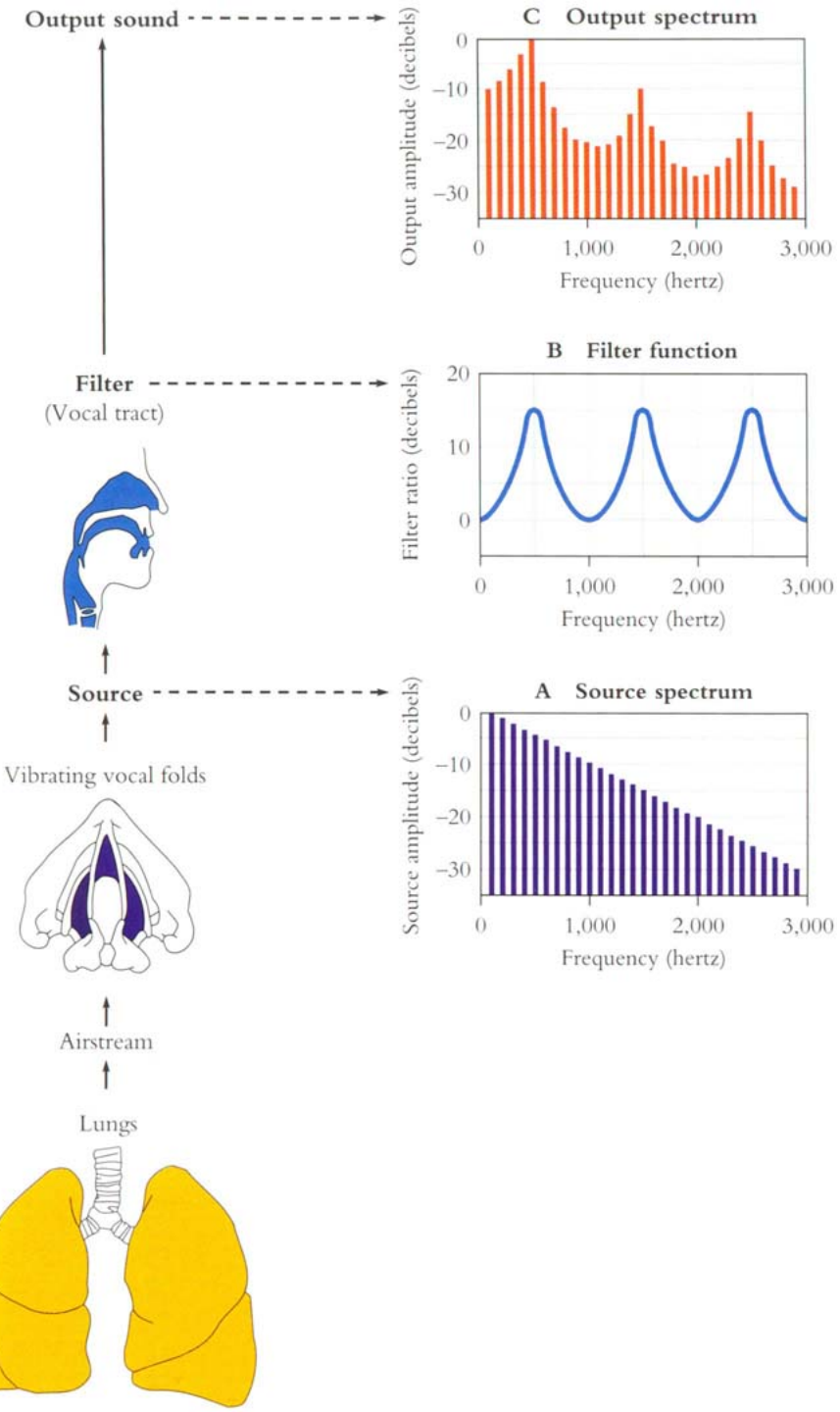


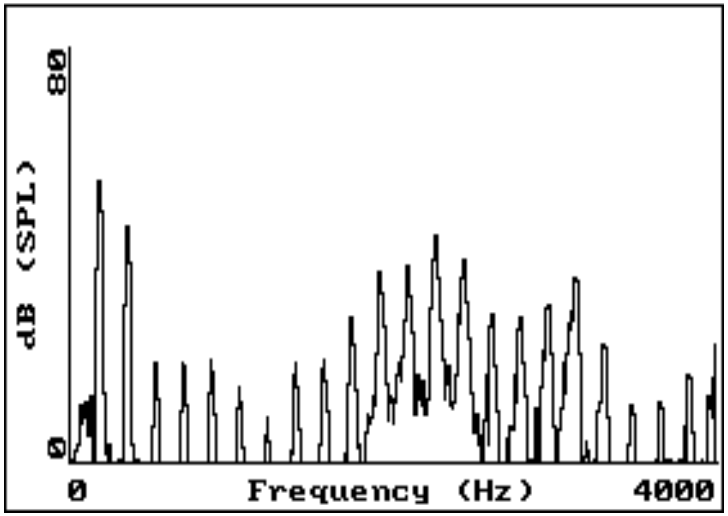
Figure: Bryan Pellom

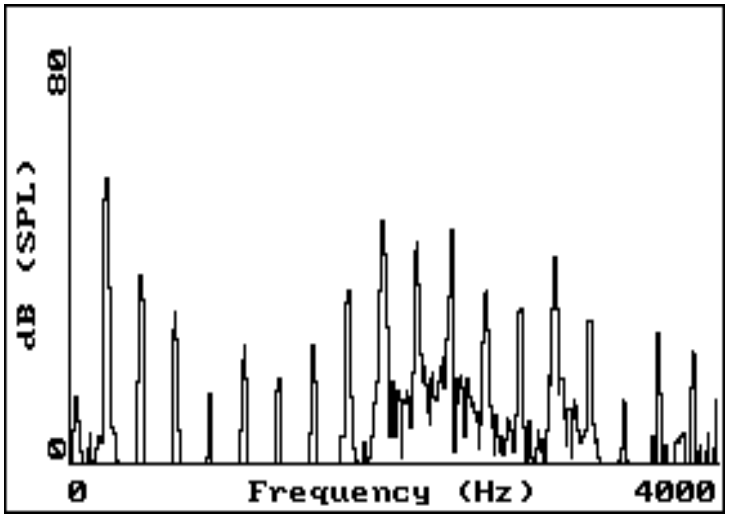
Source / Filter

- 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



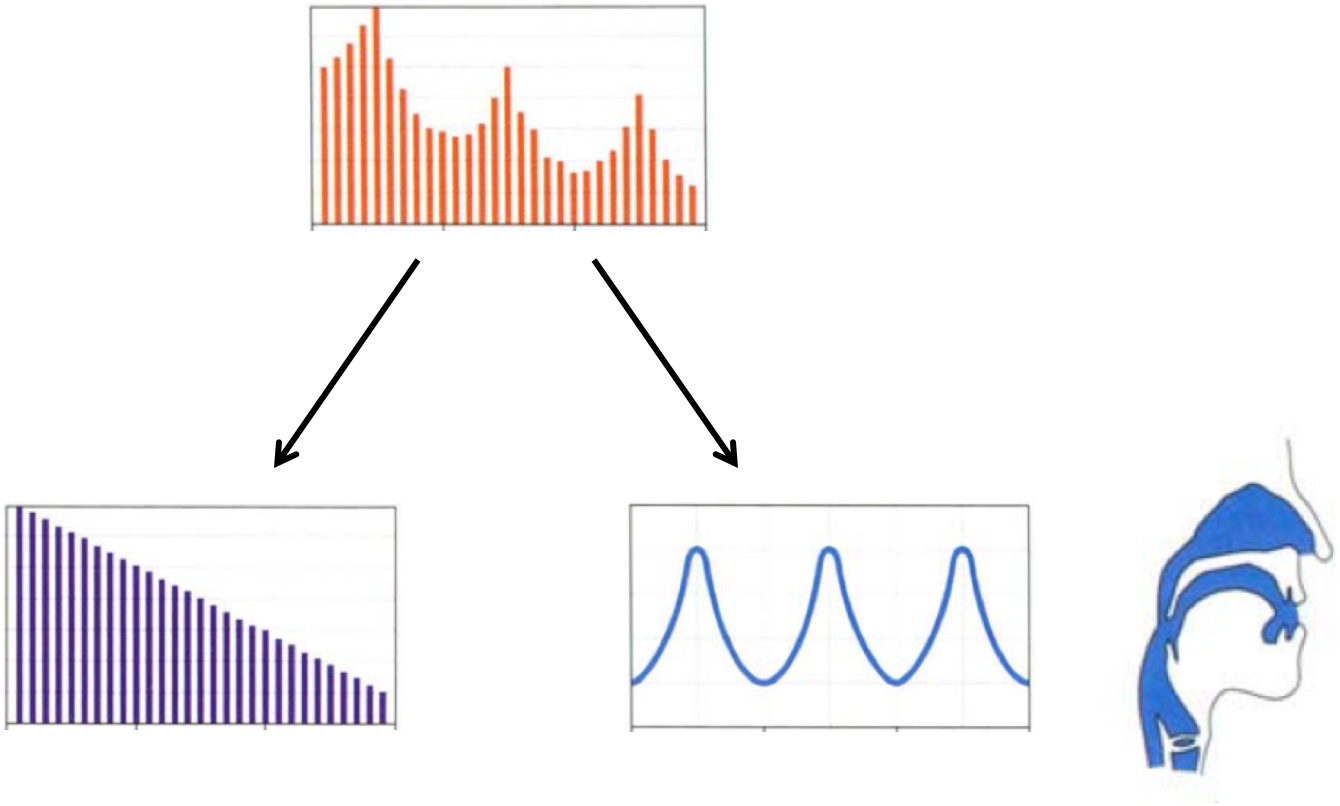
Problem with Raw Spectrum



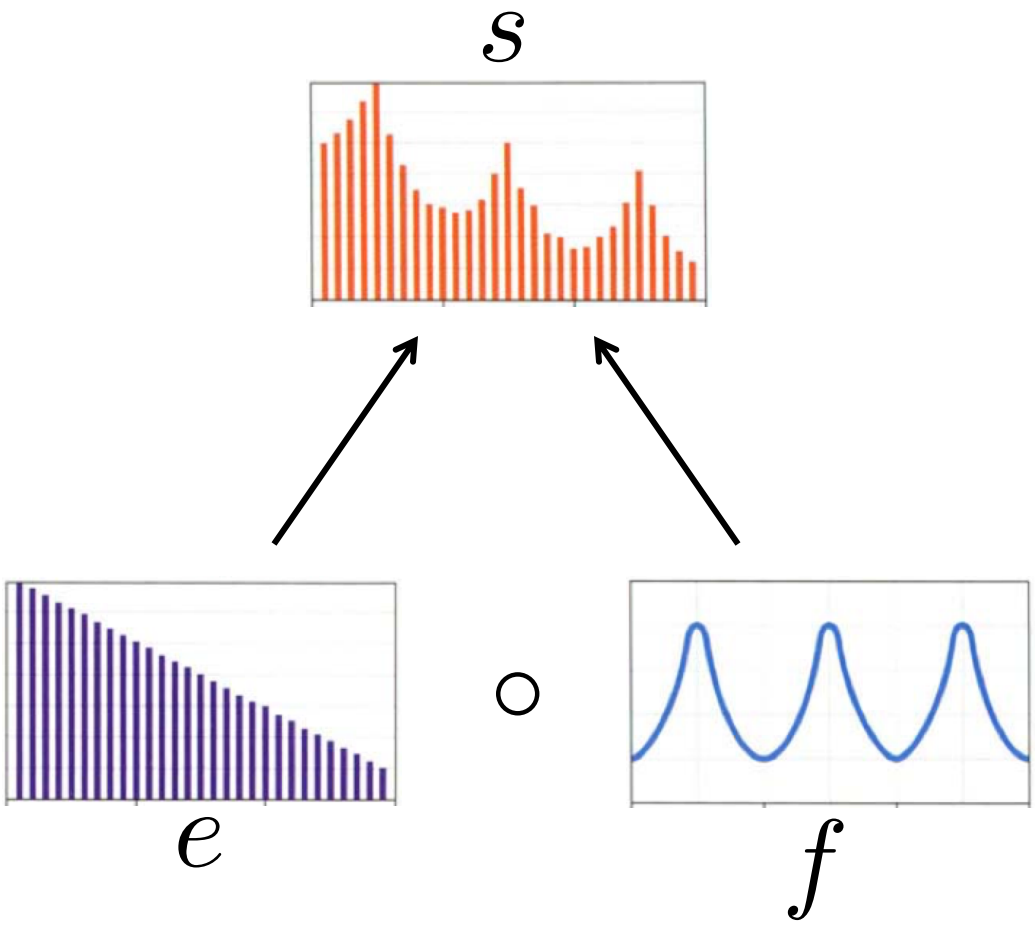


Figures from Ratree Wayland

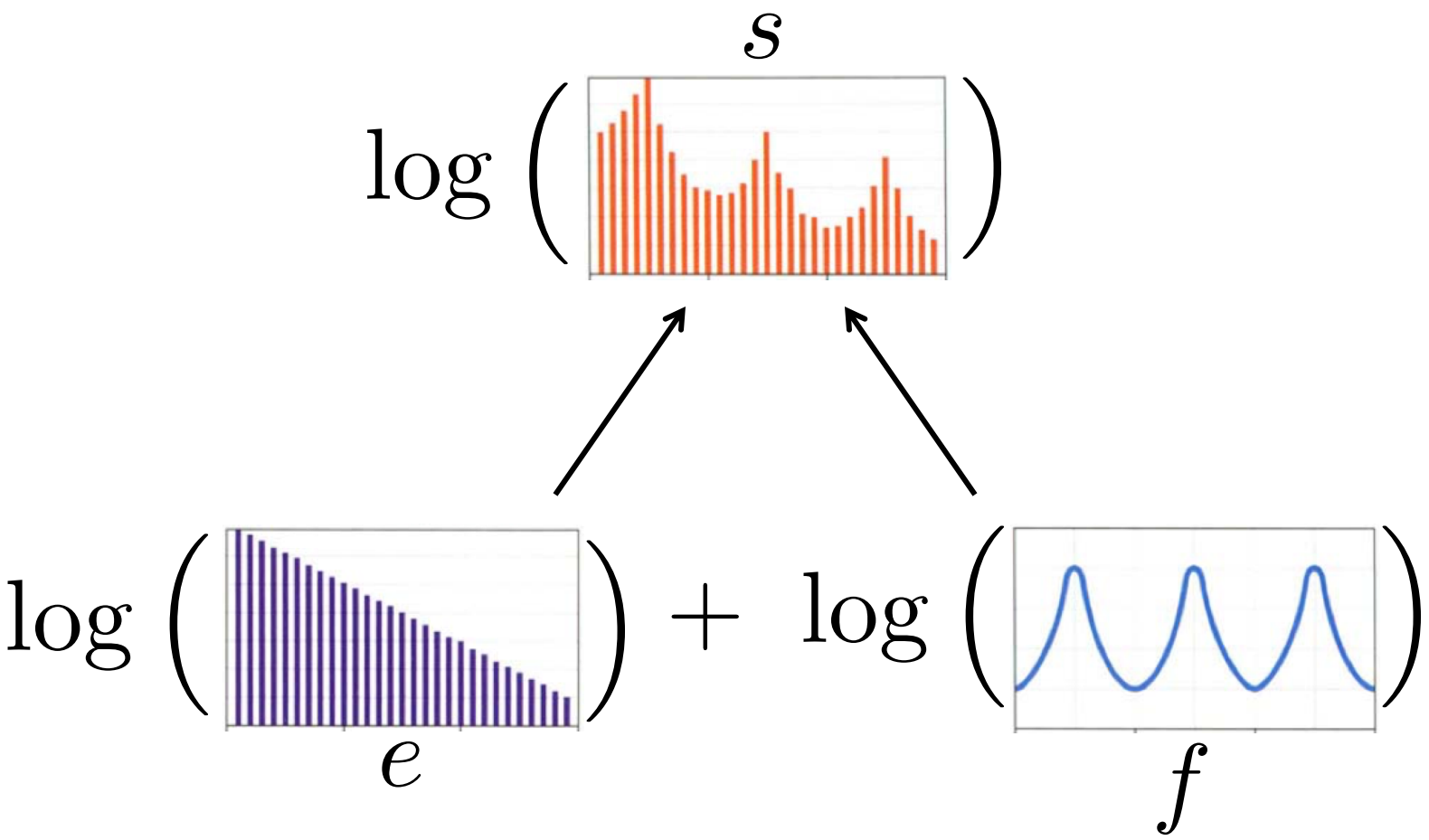
Deconvolution / Liftering



Deconvolution / Liftering



Deconvolution / Liftering

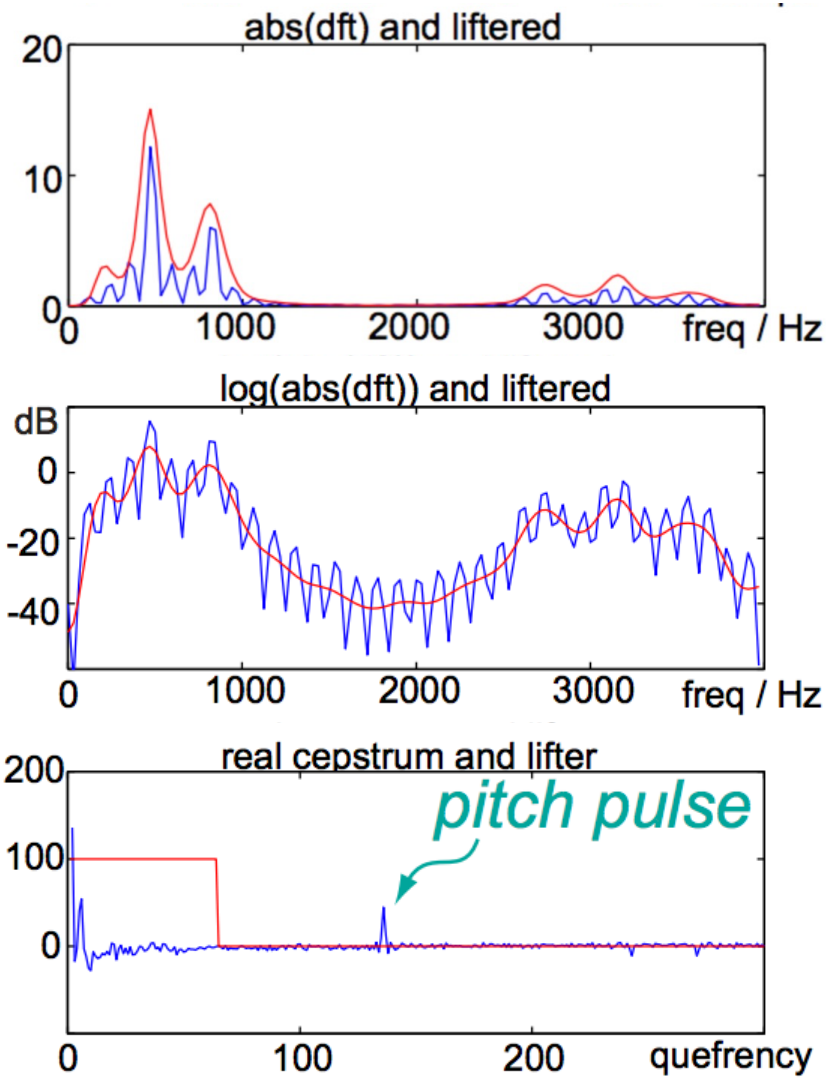


Deconvolution / Liftering

$$s = e \circ f$$

$$\log(s) = \log(e) + \log(f)$$

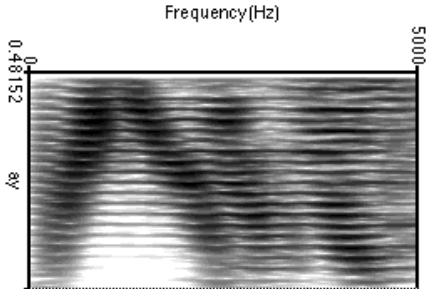
$$\mathrm{IDFT}(\log(s))$$

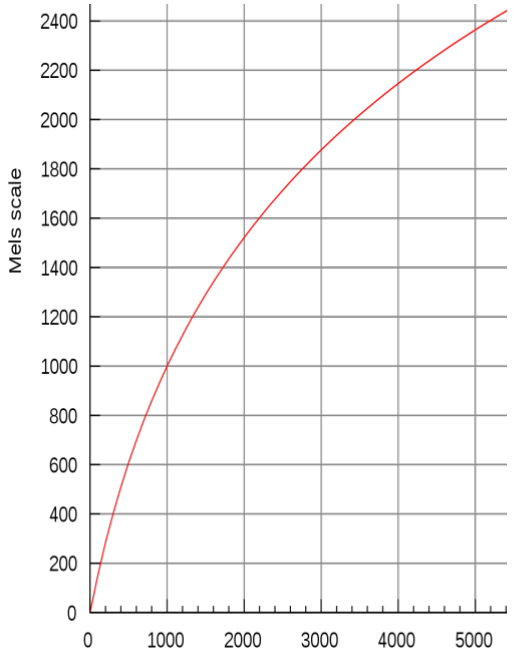


Graphs from Dan Ellis

Mel Freq. Cepstral Coefficients

- Do FFT to get spectral information
  - Like the spectrogram we saw earlier
- Apply Mel scaling (New)
  - Models human ear; more sensitivity in lower freqs
  - Approx linear below 1kHz, log above, equal samples above and below 1kHz
- Take Log
- Do discrete cosine transform





[Graph: Wikipedia]

Final Feature Vector

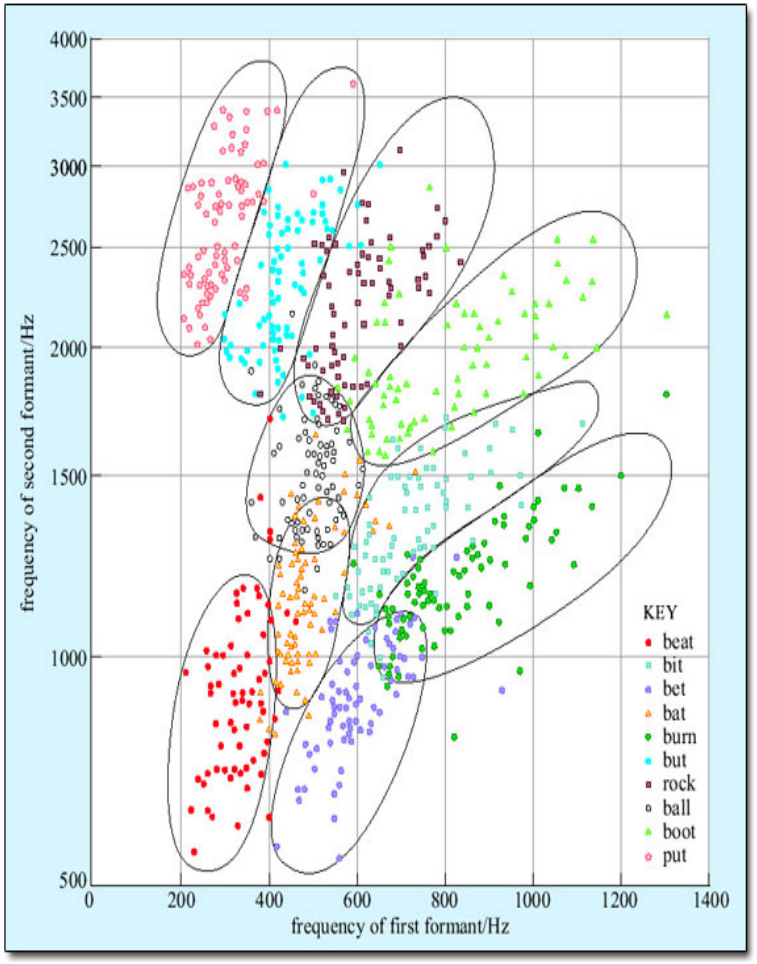
- 39 (real) features per 10 ms frame:
  - 12 MFCC features
  - 12 delta MFCC features
  - 12 delta-delta MFCC features
  - 1 (log) frame energy
  - 1 delta (log) frame energy
  - 1 delta-delta (log frame energy)

- So each frame is represented by a 39D vector

Emission Model

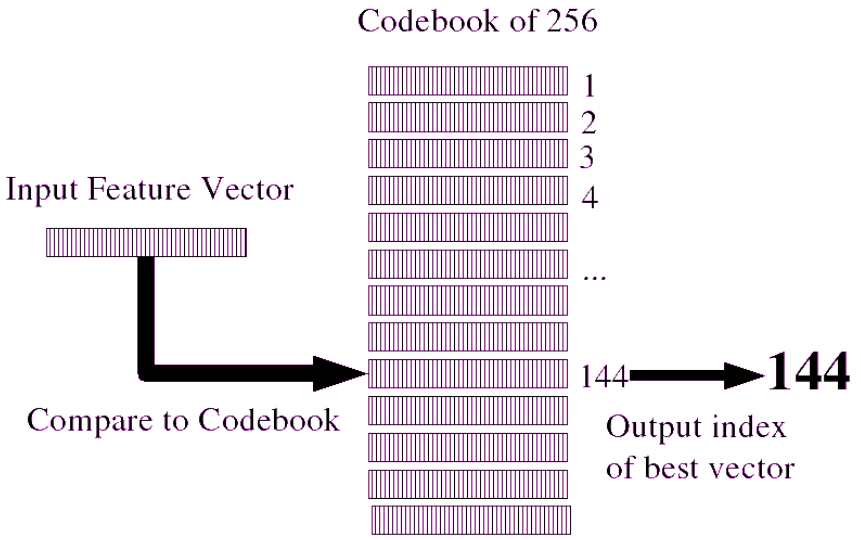
HMMs for Continuous Observations

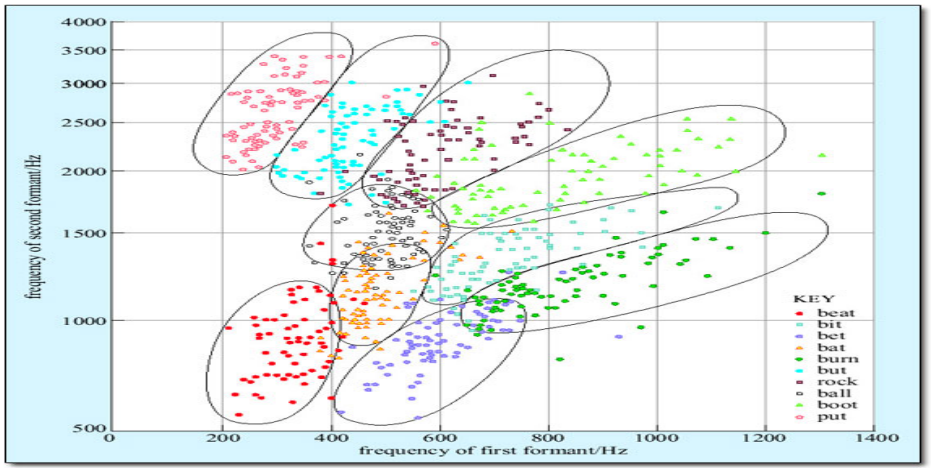
- Before: discrete set of observations
- Now: feature vectors are real-valued
- Solution 1: discretization
- Solution 2: continuous emissions
  - Gaussians
  - Multivariate Gaussians
  - Mixtures of multivariate Gaussians
- A state is progressively
  - Context independent subphone (~3 per phone)
  - Context dependent phone (triphones)
  - State tying of CD phone



Vector Quantization

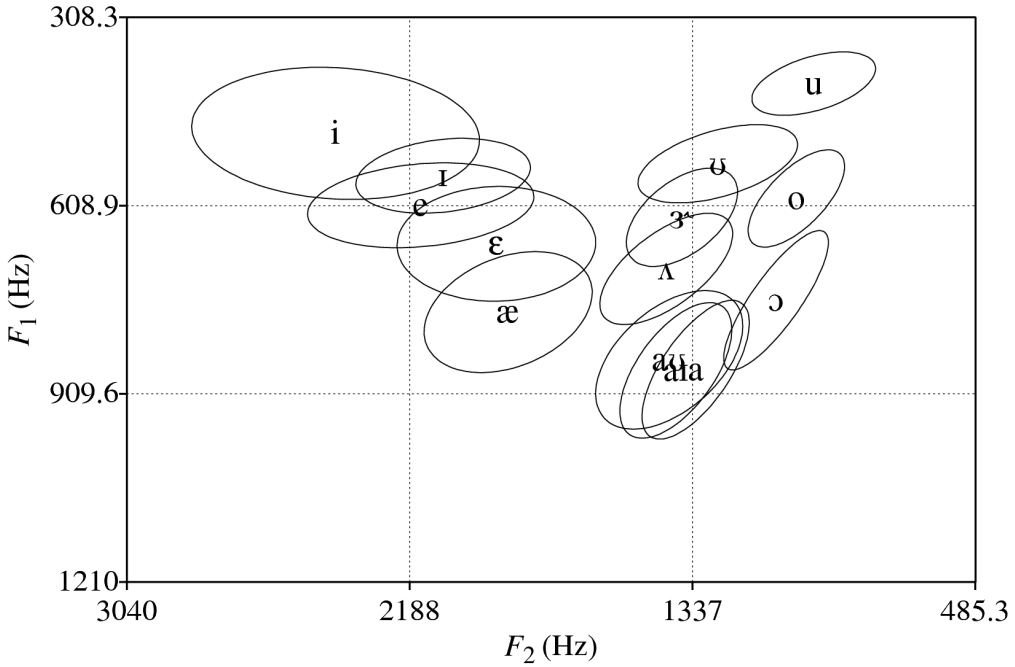
- Idea: discretization
  - Map MFCC vectors onto discrete symbols
  - Compute probabilities just by counting
- This is called vector quantization or VQ
- Not used for ASR any more
- But: useful to consider as a starting point





Gaussian Emissions

- VQ is insufficient for top-quality ASR
  - Hard to cover high-dimensional space with codebook
  - Moves ambiguity from the model to the preprocessing
- Instead: assume the possible values of the observation vectors are normally distributed.
  - Represent the observation likelihood function as a Gaussian?



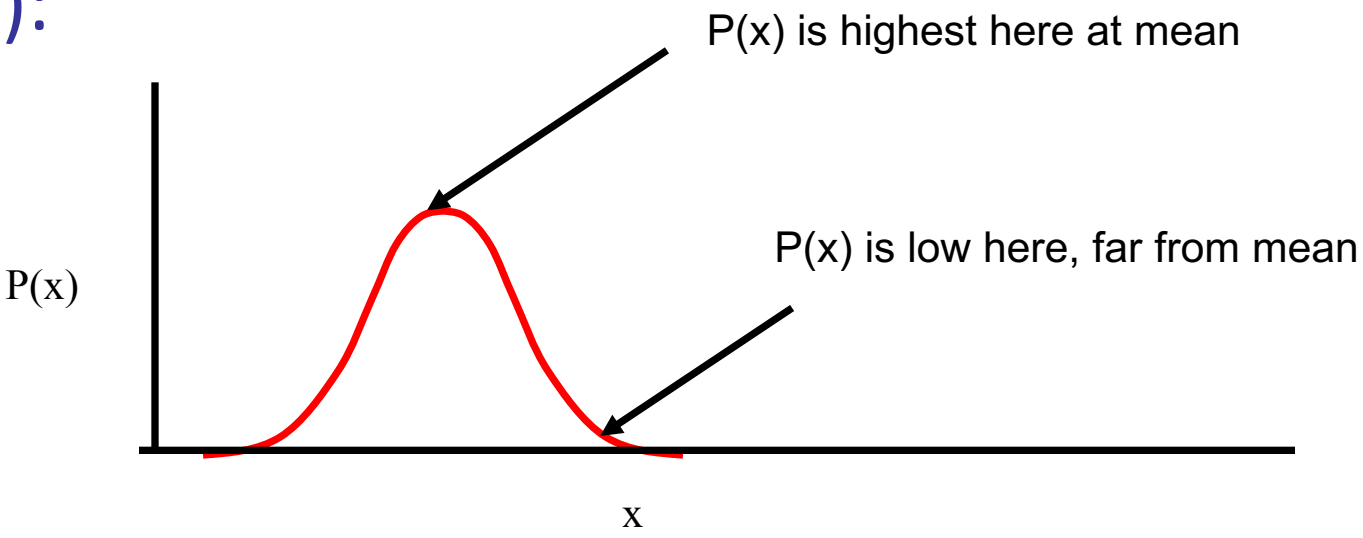
From bartus.org/akustyk

Gaussians for Acoustic Modeling

A Gaussian is parameterized by a mean and a variance:

$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- P(x):



Multivariate Gaussians

- Instead of a single mean μ and variance σ^2:

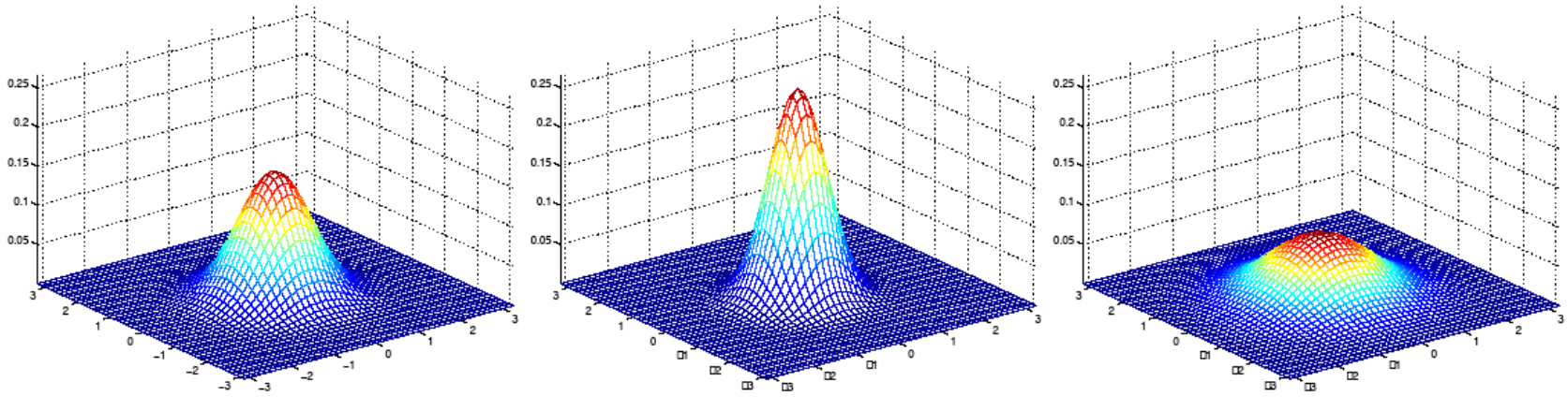
$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Vector of means μ and covariance matrix Σ

$$P(x|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

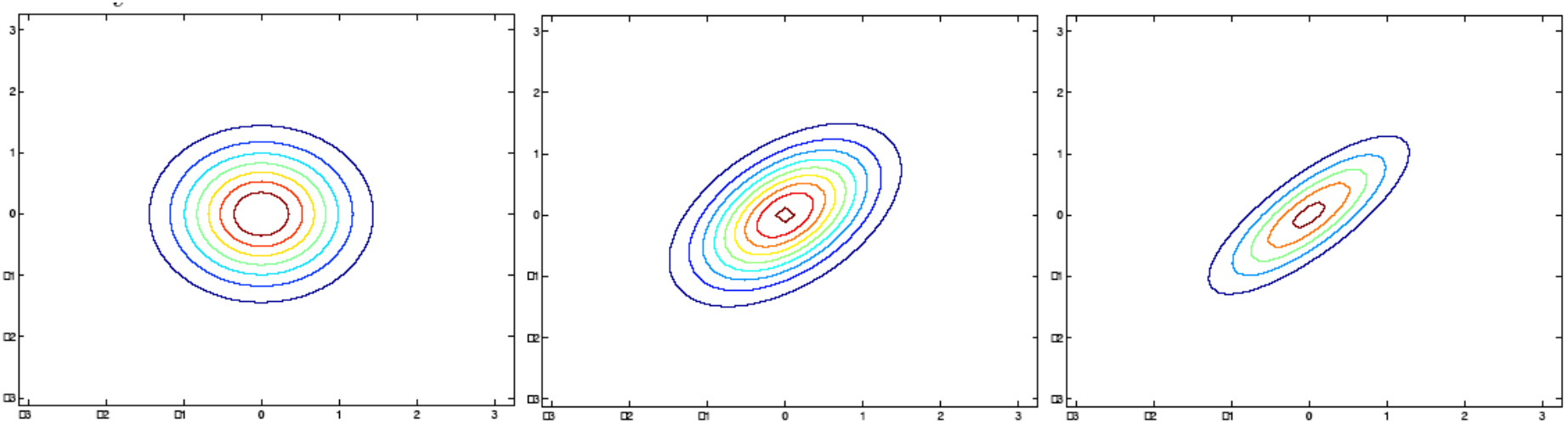
- Usually assume diagonal covariance (!)
  - This isn't very true for FFT features, but is less bad for MFCC features

Gaussians: Size of Σ



- $\mu = [0\ 0]$ $\mu = [0\ 0]$ $\mu = [0\ 0]$
- $\Sigma = I$ $\Sigma = 0.6I$ $\Sigma = 2I$
- As Σ becomes larger, Gaussian becomes more spread out; as Σ becomes smaller, Gaussian more compressed

Text and figures from Andrew Ng



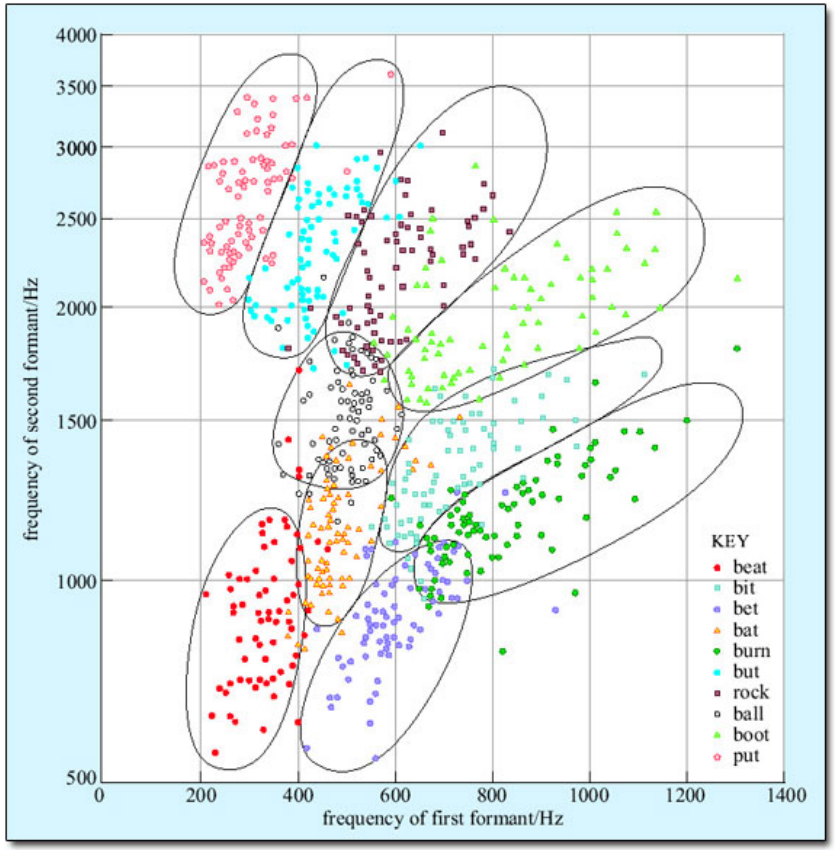
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad .\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

- As we increase the off diagonal entries, more correlation between value of x and value of y

Text and figures from Andrew Ng

But we're not there yet

- Single Gaussians may do a bad job of modeling a complex distribution in any dimension
- Even worse for diagonal covariances
- Solution: mixtures of Gaussians



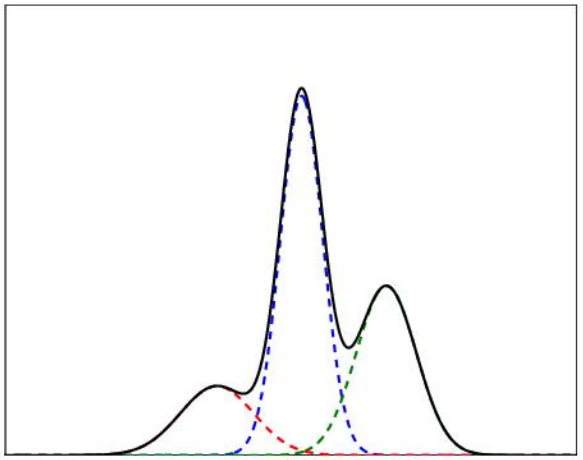
From openlearn.open.ac.uk

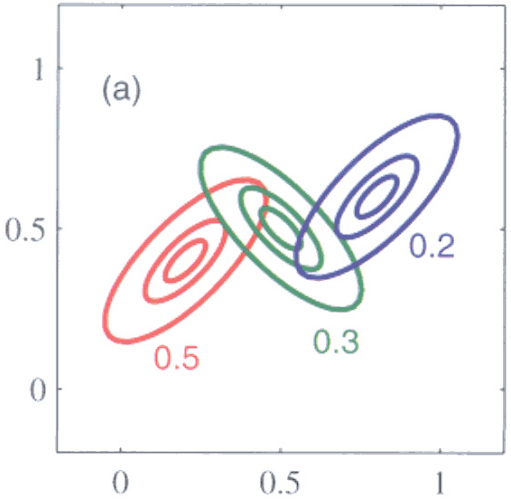
Mixtures of Gaussians

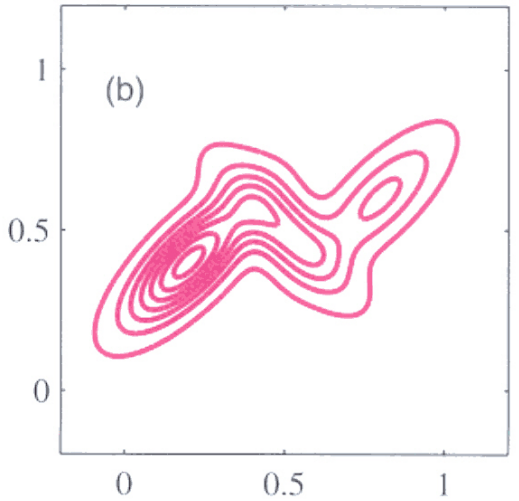
- Mixtures of Gaussians:

$$P(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{k/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i)\right)$$

$$P(x|\mu, \mathbf{\Sigma}, \mathbf{c}) = \sum_i c_i P(x|\mu_i, \Sigma_i)$$



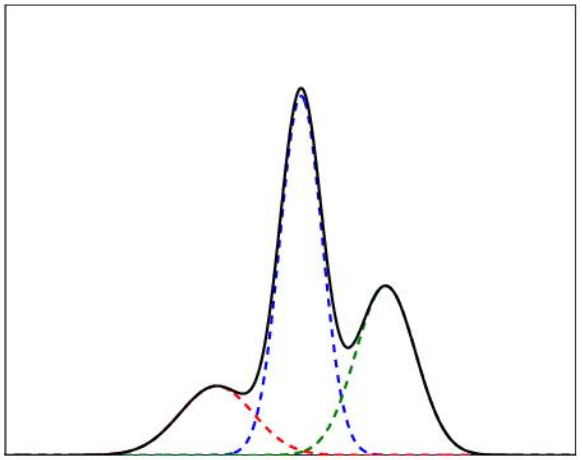




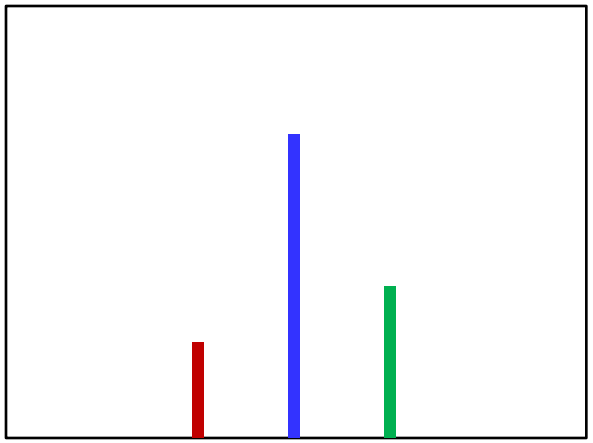
From robots.ox.ac.uk

GMMs

- Summary: each state has an emission distribution P(x|s) (likelihood function) parameterized by:
  - M mixture weights
  - M mean vectors of dimensionality D
  - Either M covariance matrices of DxD or M Dx1 diagonal variance vectors



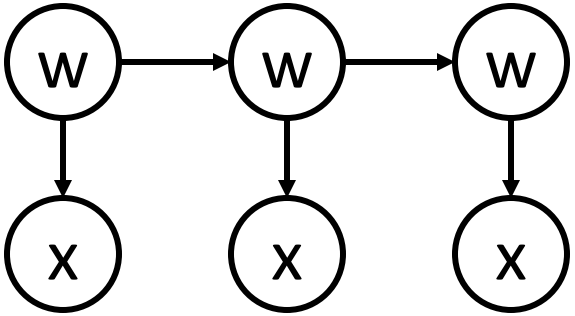
- Like soft vector quantization after all
  - Think of the mixture means as being learned codebook entries
  - Think of the Gaussian densities as a learned codebook distance function
  - Think of the mixture of Gaussians like a multinomial over codes
  - (Even more true given shared Gaussian inventories, cf next week)



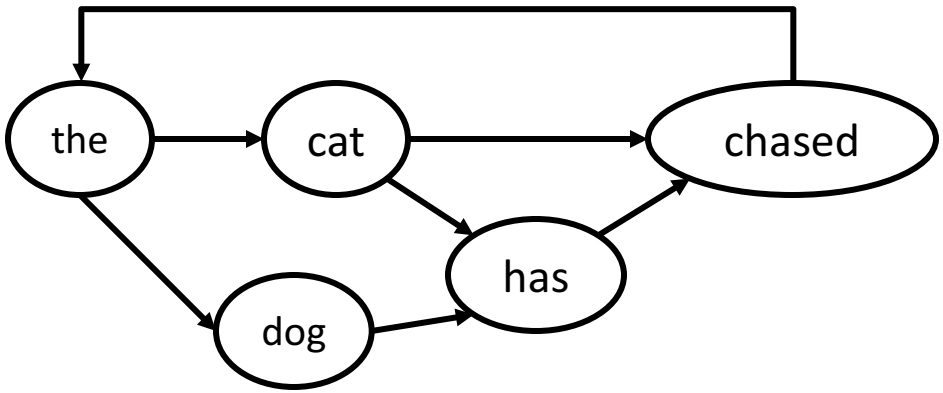
State Model

State Transition Diagrams

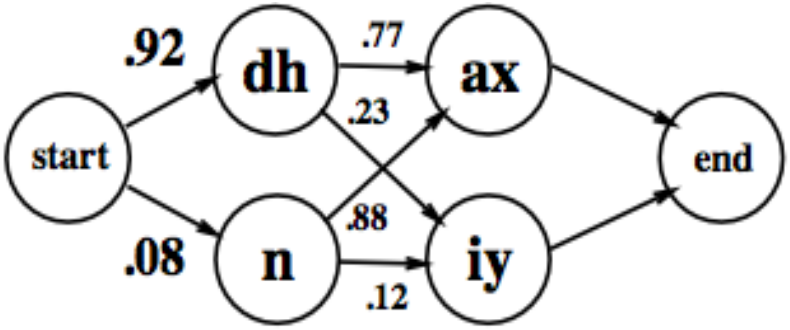
- Bayes Net: HMM as a Graphical Model



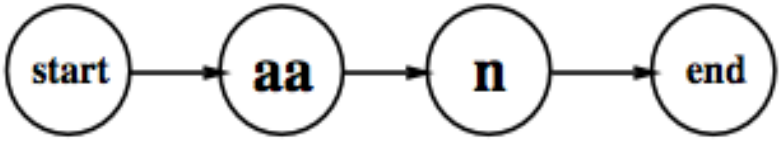
- State Transition Diagram: Markov Model as a Weighted FSA



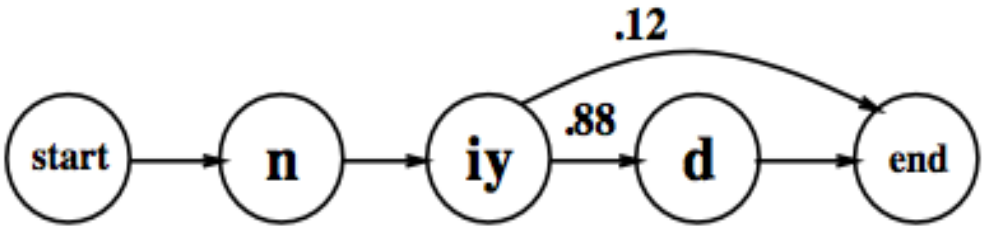
ASR Lexicon



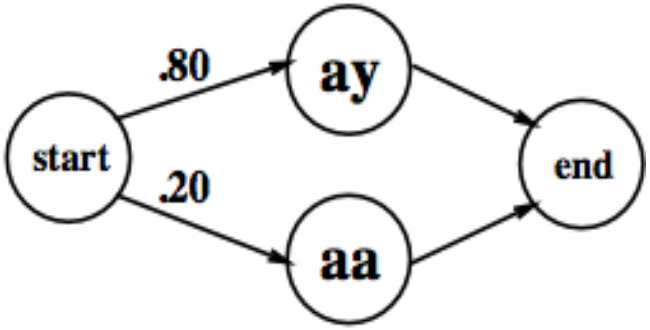
Word model for "the"



Word model for "on"



Word model for "need"



Word model for "I"

Figure: J & M

Lexical State Structure

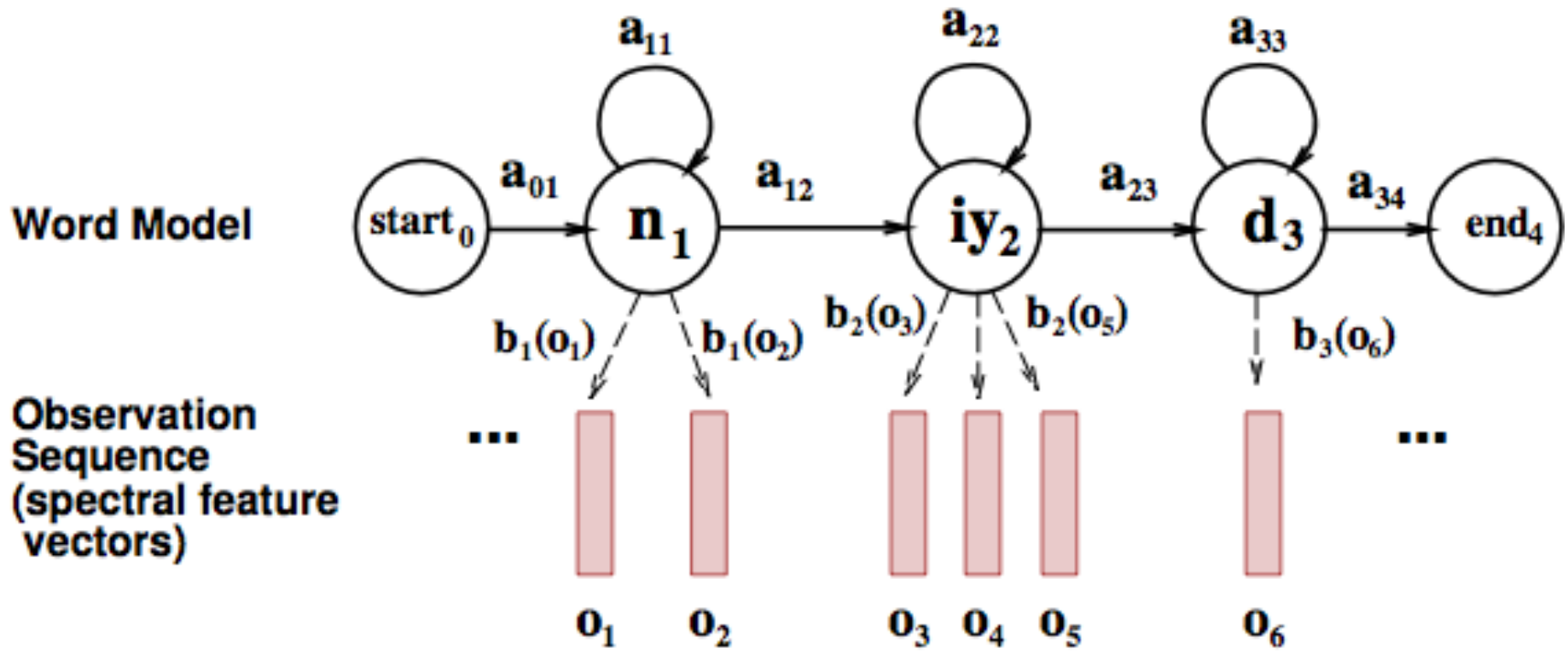


Figure: J & M

Adding an LM

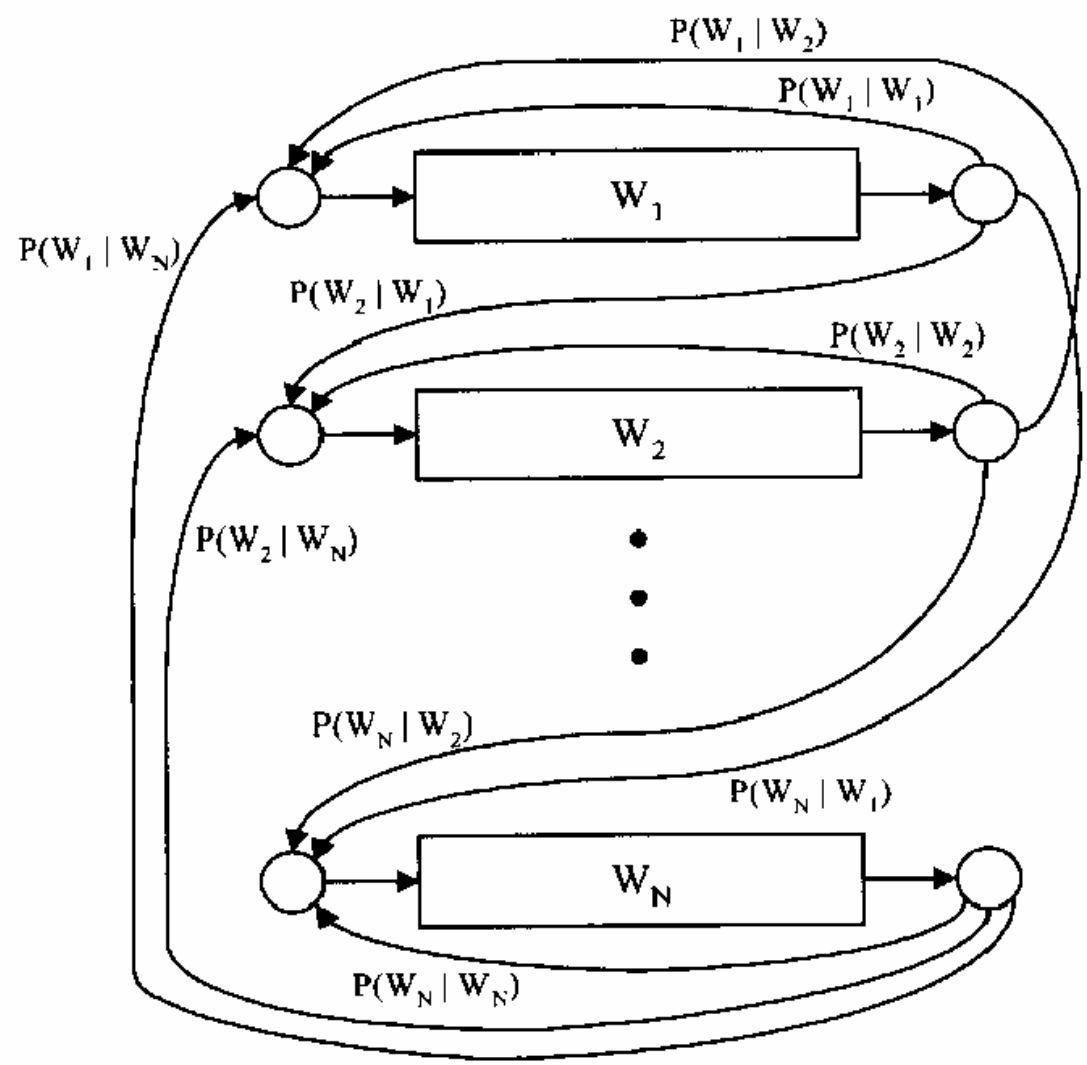


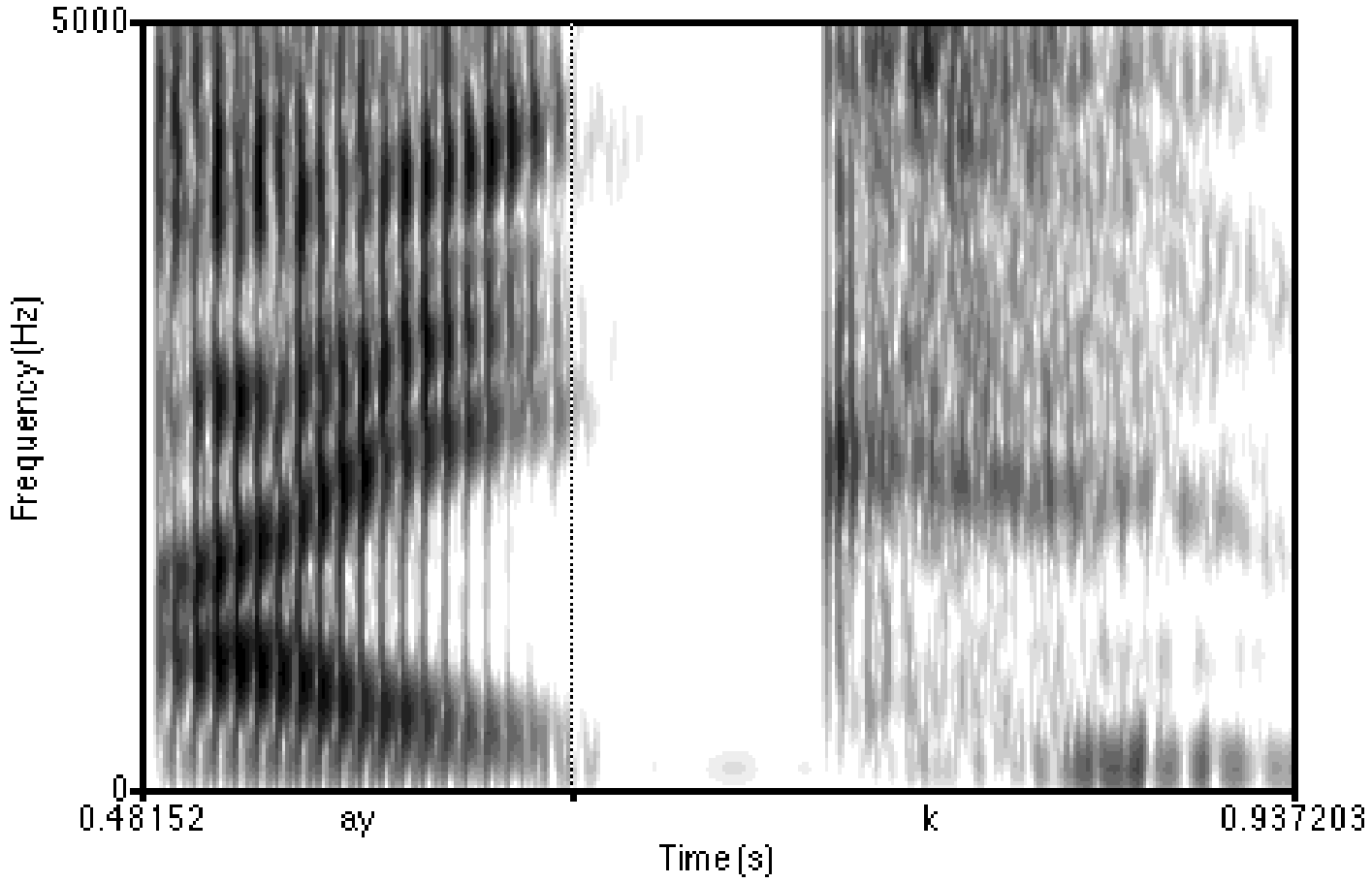
Figure from Huang et al page 618

State Space

- State space must include
  - Current word (|V| on order of 20K+)
  - Index within current word (|L| on order of 5)
  - E.g. (lec[t]ure) (though not in orthography!)
- Acoustic probabilities only depend on phone type
  - E.g. P(x|lec[t]ure) = P(x|t)
- From a state sequence, can read a word sequence

State Refinement

Phones Aren't Homogeneous



Need to Use Subphones

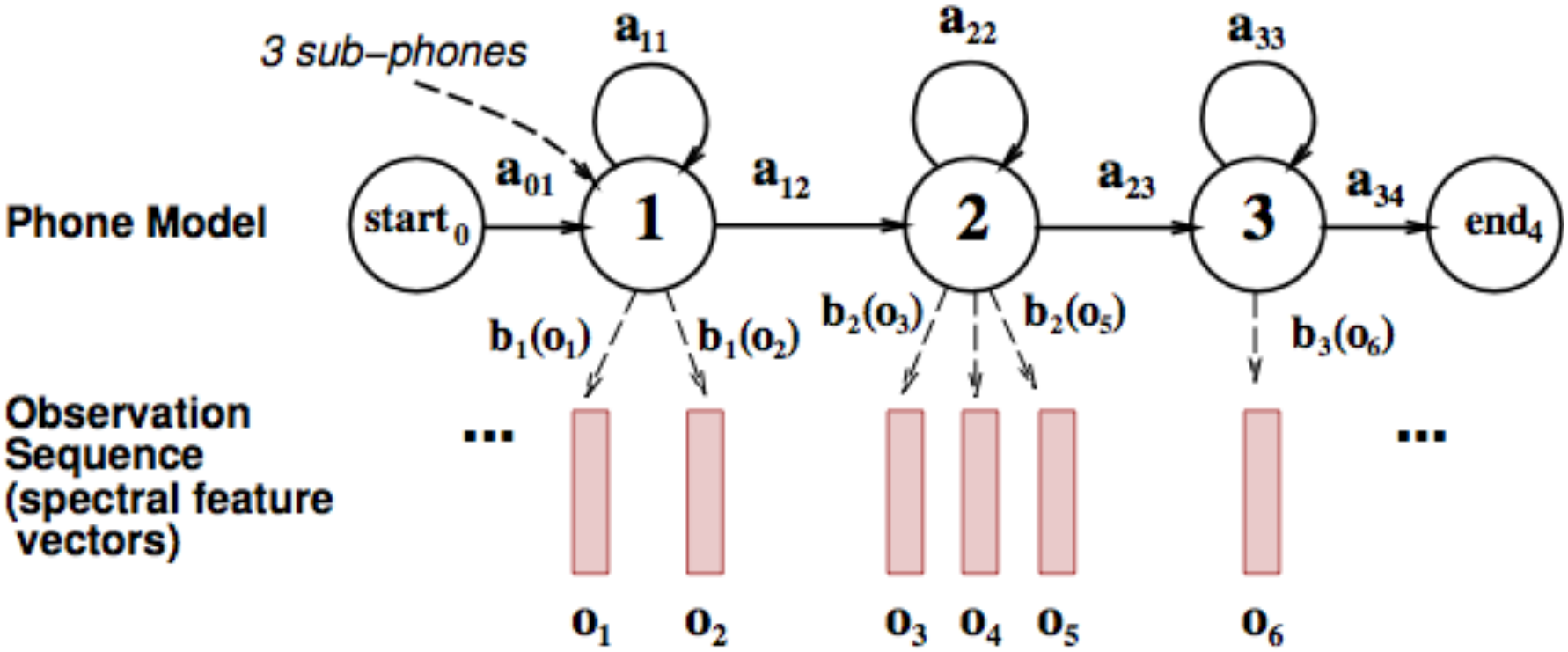


Figure: J & M

A Word with Subphones

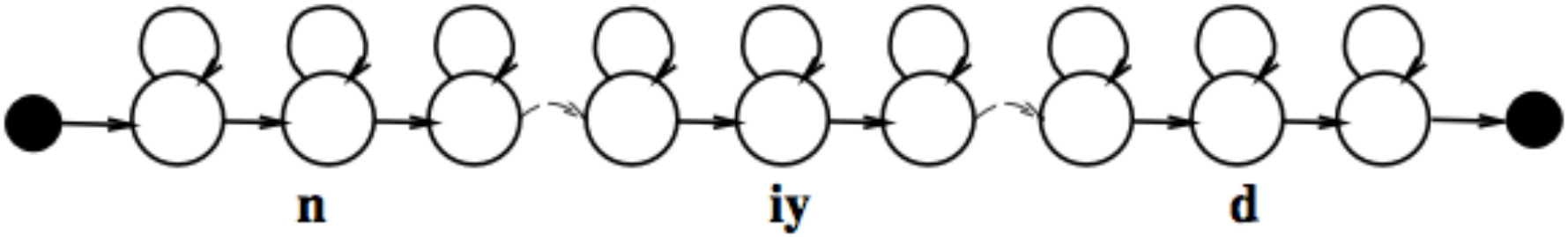
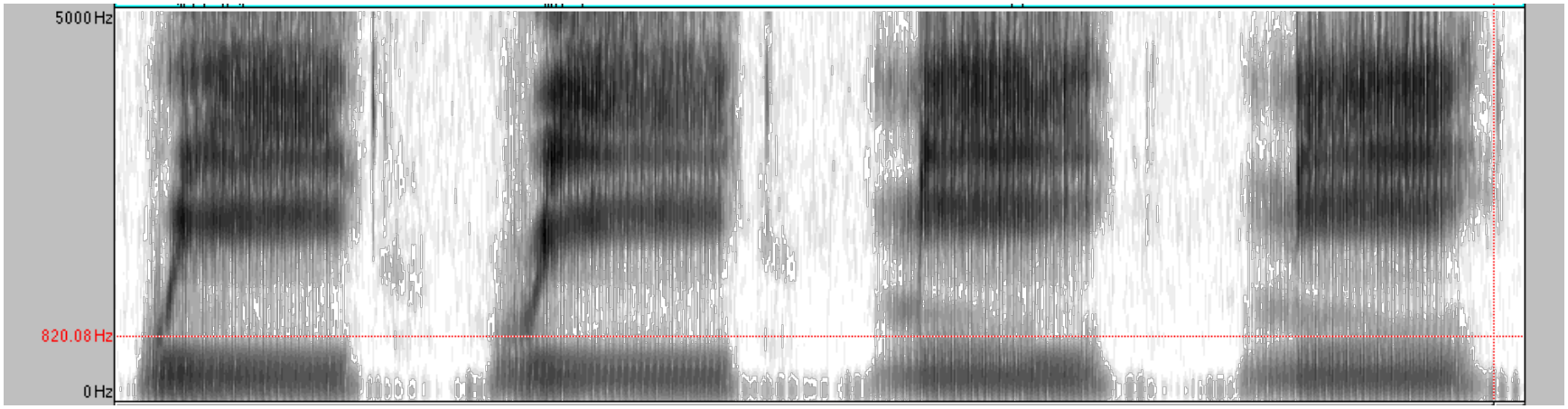


Figure: J & M

Modeling phonetic context



w iy r iy m iy n iy

“Need” with triphone models

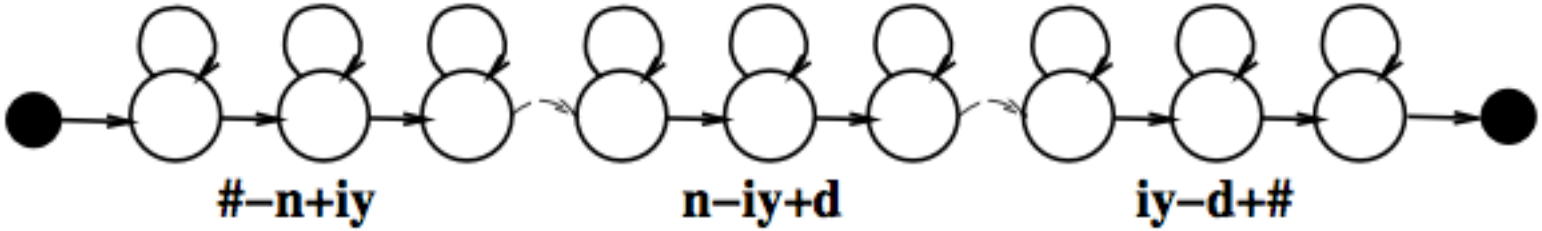


Figure: J & M

Lots of Triphones

- Possible triphones: 50x50x50=125,000
- How many triphone types actually occur?
- 20K word WSJ Task (from Bryan Pellom)
  - Word internal models: need 14,300 triphones
  - Cross word models: need 54,400 triphones
- Need to generalize models, tie triphones

State Tying / Clustering

- [Young, Odell, Woodland 1994]
- How do we decide which triphones to cluster together?
- Use phonetic features (or 'broad phonetic classes')
  - Stop
  - Nasal
  - Fricative
  - Sibilant
  - Vowel
  - lateral

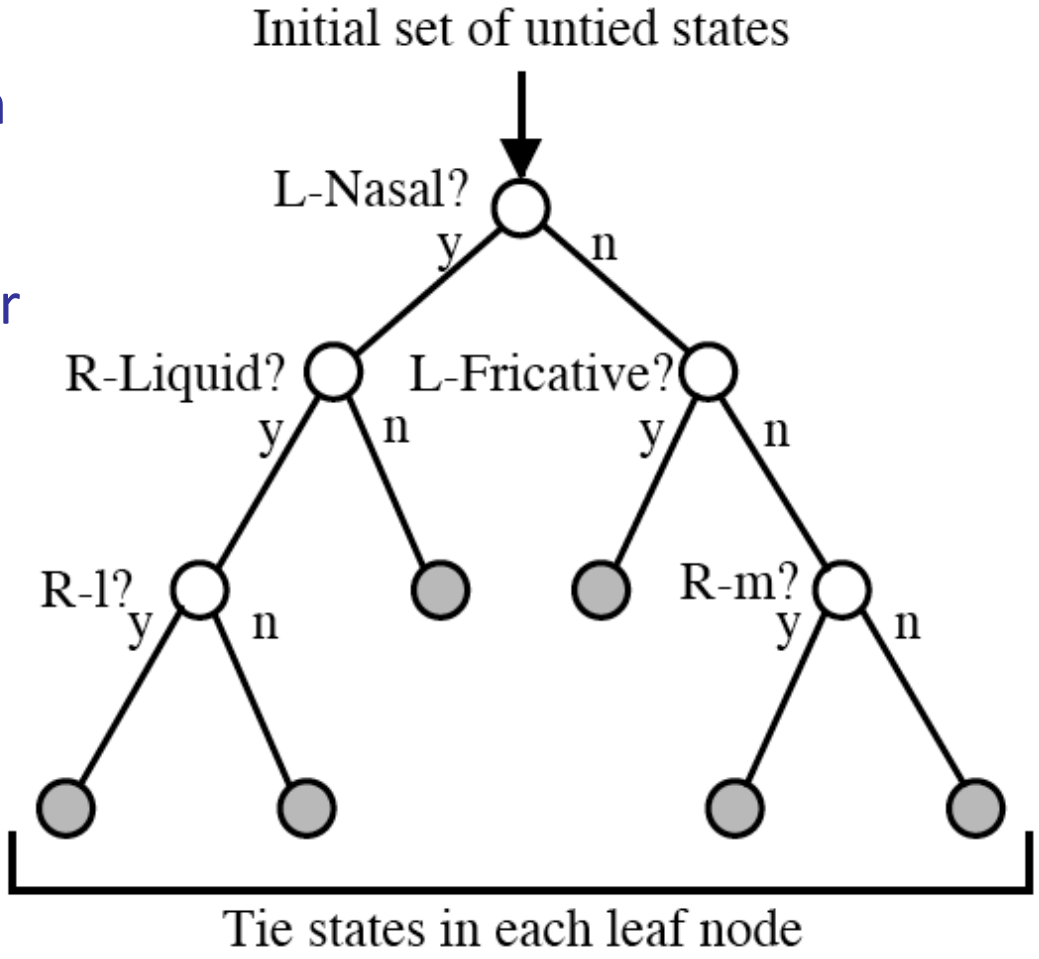


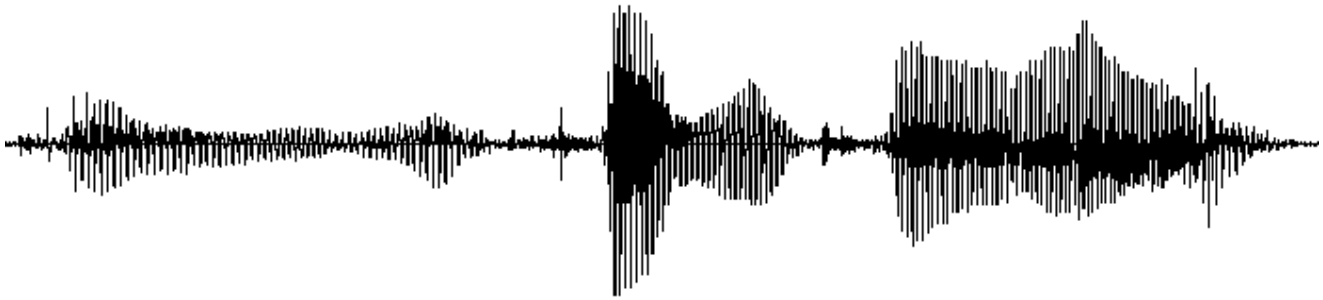
Figure: J & M

State Space

- State space now includes
  - Current word: |W| is order 20K
  - Index in current word: |L| is order 5
  - Subphone position: 3
  - E.g. (lec[t-mid]ure)
- Acoustic model depends on clustered phone context
  - But this doesn't grow the state space
- But, adding the LM context for trigram+ does
  - (after the, lec[t-mid]ure)
  - This is a real problem for decoding

Decoding

Inference Tasks



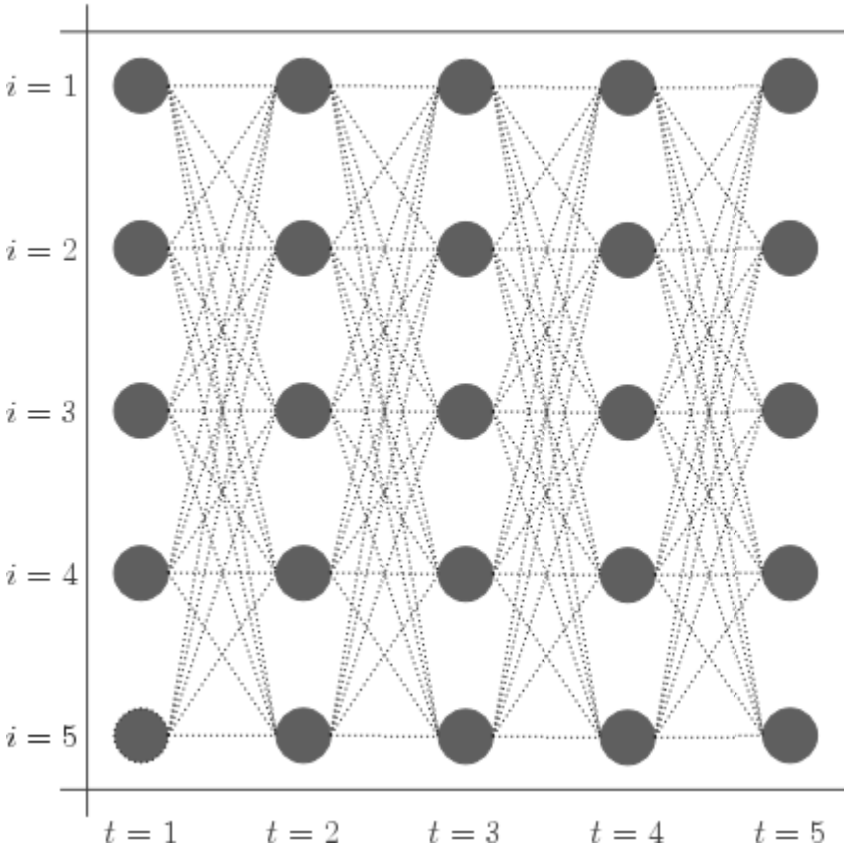
Most likely word sequence:

d - ae - d

Most likely state sequence:

d_1-d_6-d_6-d_4-ae_5-ae_2-ae_3-ae_0-d_2-d_2-d_3-d_7-d_5

Viterbi Decoding



$$\phi_t(s_t, s_{t-1}) = P(x_t|s_t)P(s_t|s_{t-1})$$

$$v_t(s_t) = \max_{s_{t-1}} \phi_t(s_t, s_{t-1}) v_{t-1}(s_{t-1})$$

Figure: Enrique Benimeli

Viterbi Decoding

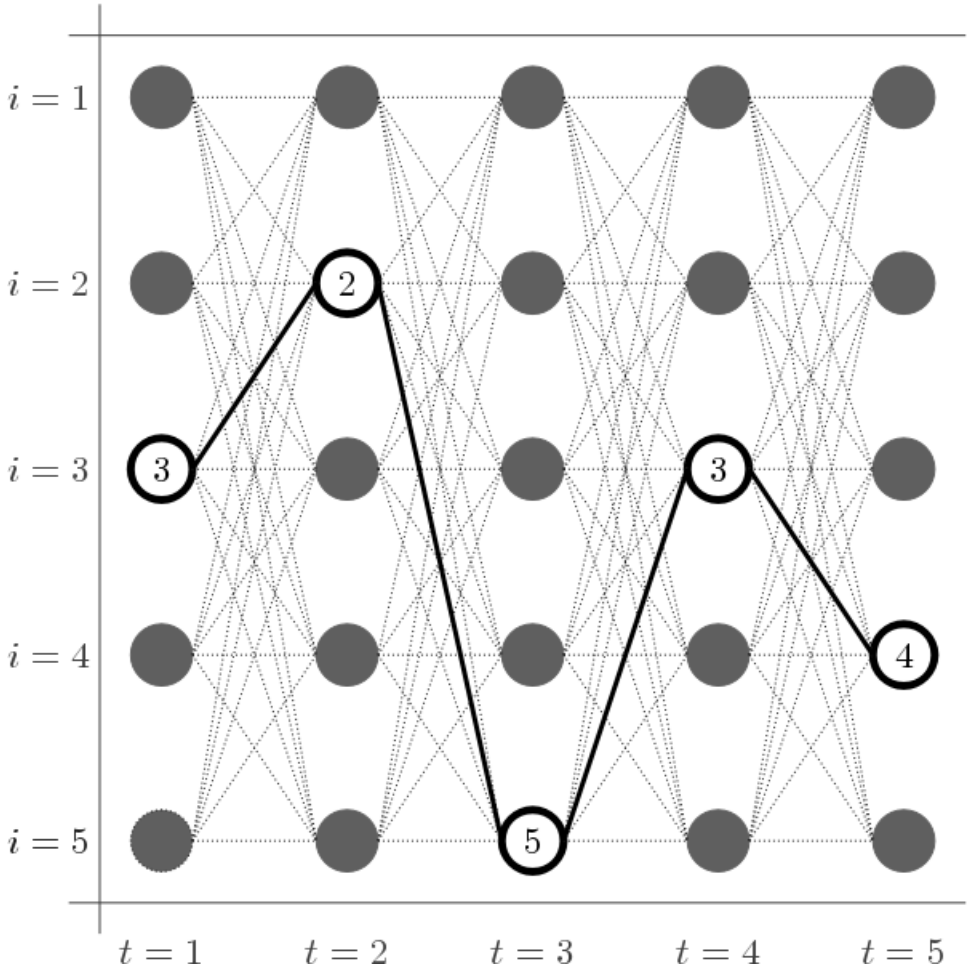
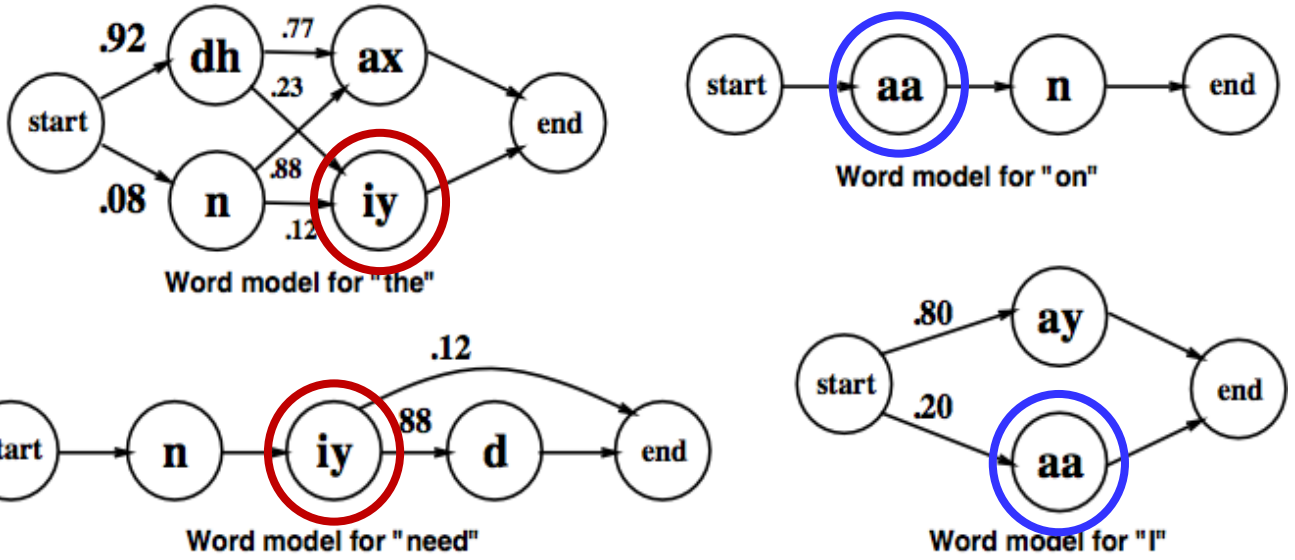


Figure: Enrique Benimeli

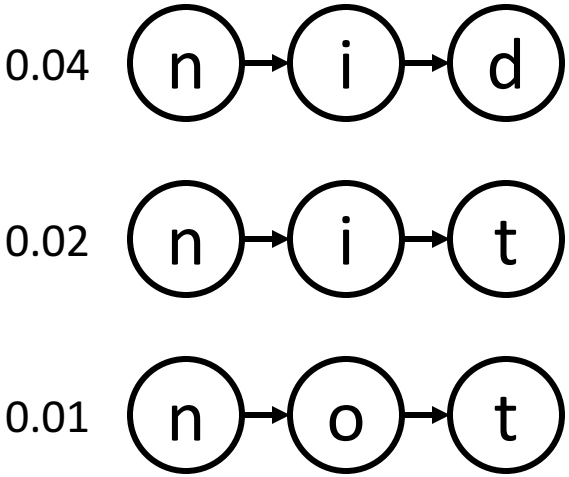
Emission Caching

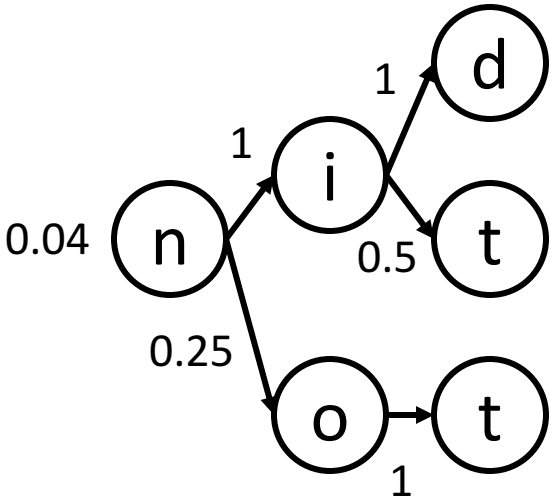
- Problem: scoring all the P(x|s) values is too slow
- Idea: many states share tied emission models, so cache them



Prefix Trie Encodings

- Problem: many partial-word states are indistinguishable
- Solution: encode word production as a prefix trie (with pushed weights)



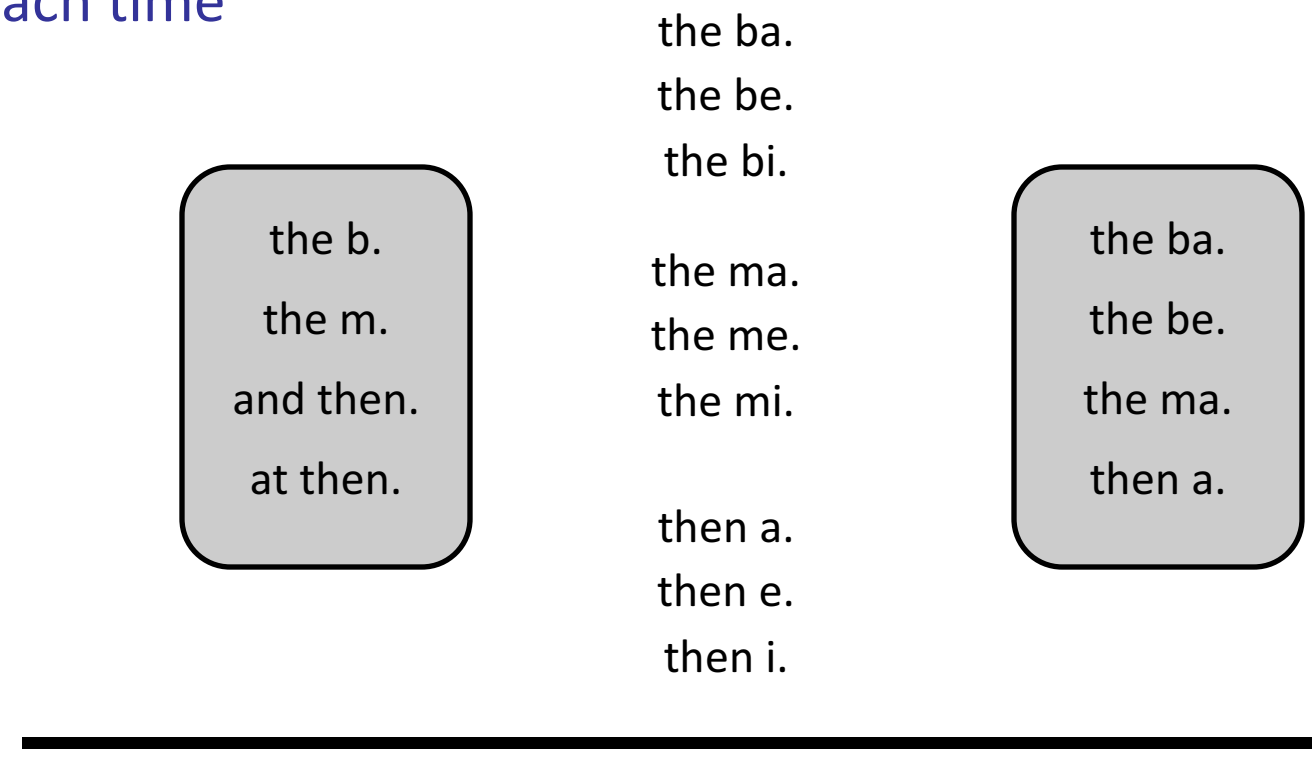


- A specific instance of minimizing weighted FSAs [Mohri, 94]

Figure: Aubert, 02

Beam Search

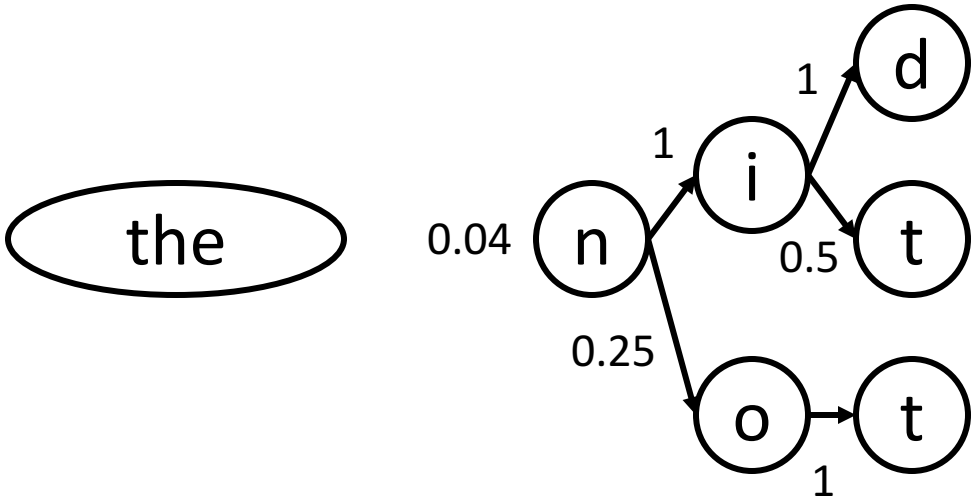
- Problem: trellis is too big to compute v(s) vectors
- Idea: most states are terrible, keep v(s) only for top states at each time



- Important: still dynamic programming; collapse equiv states

LM Factoring

- Problem: Higher-order n-grams explode the state space
- (One) Solution:
  - Factor state space into (word index, lm history)
  - Score unigram prefix costs while inside a word
  - Subtract unigram cost and add trigram cost once word is complete



LM Reweighting

- Noisy channel suggests

$$P(x|w)P(w)$$

- In practice, want to boost LM

$$P(x|w)P(w)^{\alpha}$$

- Also, good to have a “word bonus” to offset LM costs

$$P(x|w)P(w)^{\alpha}|w|^{\beta}$$

- These are both consequences of broken independence assumptions in the model

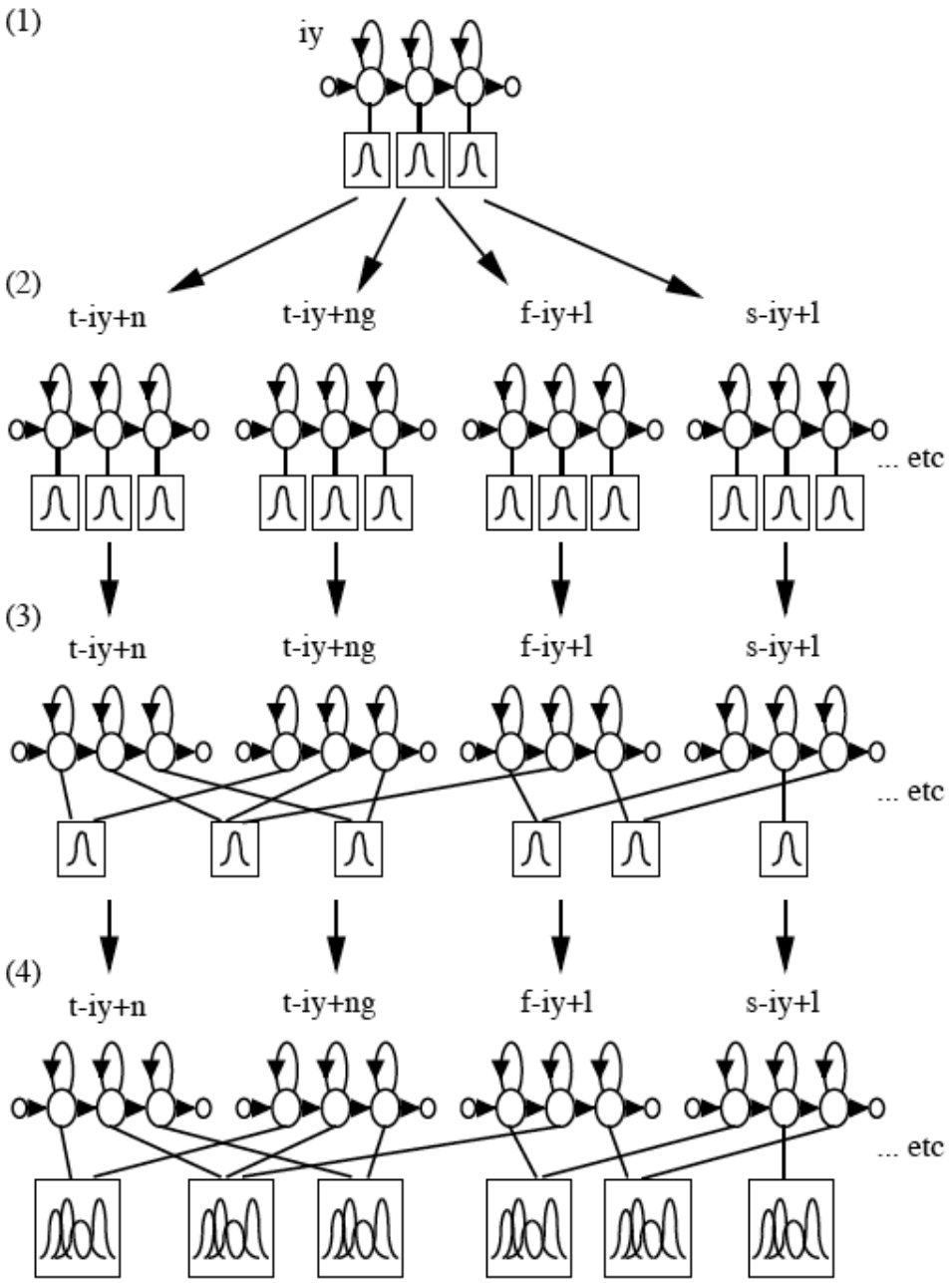
Training

Training Mixture Models

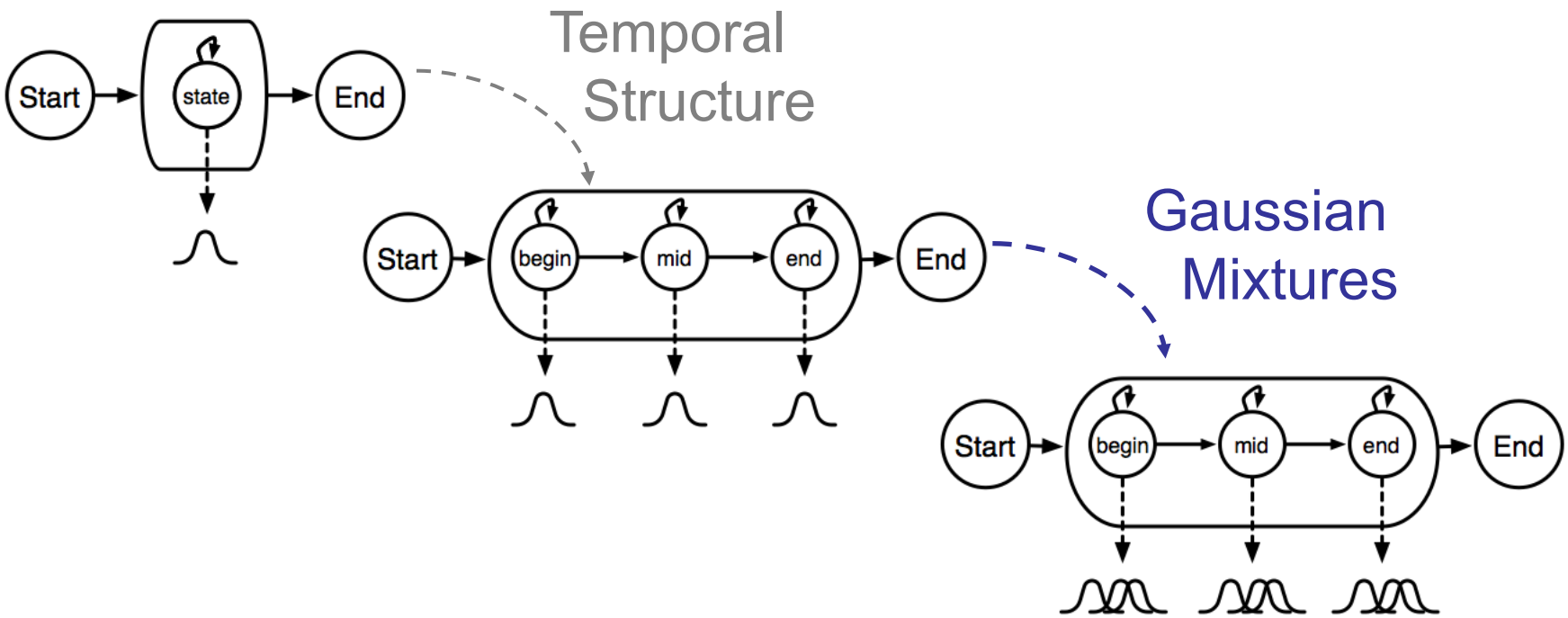
- Input: wav files with unaligned transcriptions
- Forced alignment
  - Computing the "Viterbi path" over the training data (where the transcription is known) is called "forced alignment"
  - We know which word string to assign to each observation sequence.
  - We just don't know the state sequence.
  - So we constrain the path to go through the correct words (by using a special example-specific language model)
  - And otherwise run the Viterbi algorithm
- Result: aligned state sequence

State Tying

- Creating CD phones:
  - Start with monophone, do EM training
  - Clone Gaussians into triphones
  - Build decision tree and cluster Gaussians
  - Clone and train mixtures (GMMs)
- General idea:
  - Introduce complexity gradually
  - Interleave constraint with flexibility



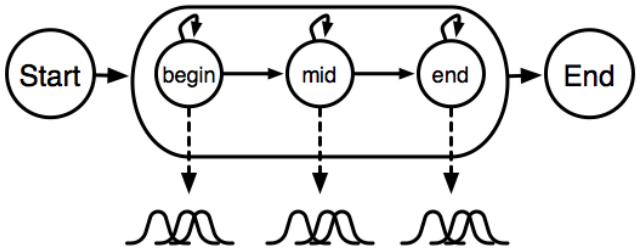
Standard subphone/mixture HMM



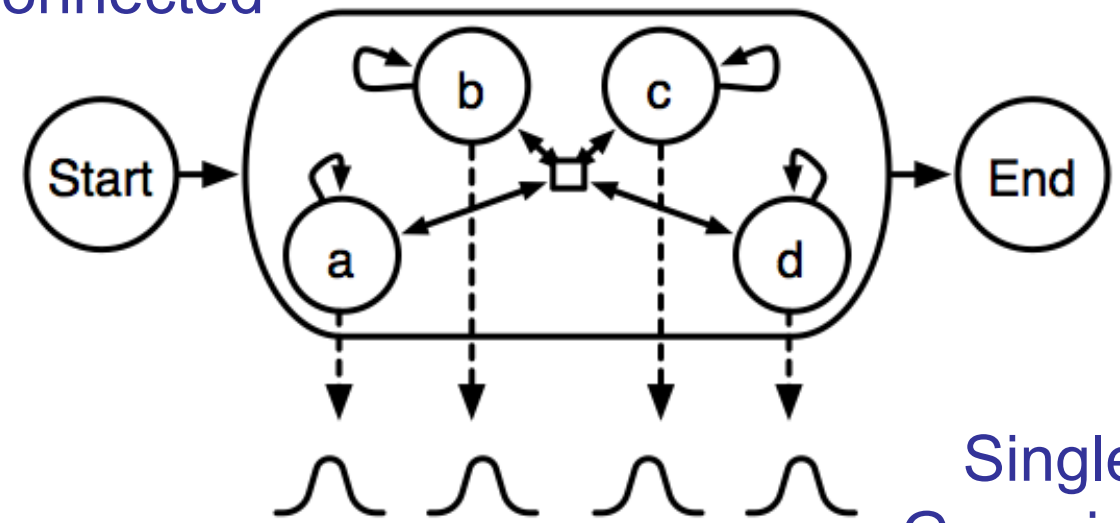
Model	Error rate
HMM Baseline	**25.1%**

An Induced Model

Standard Model



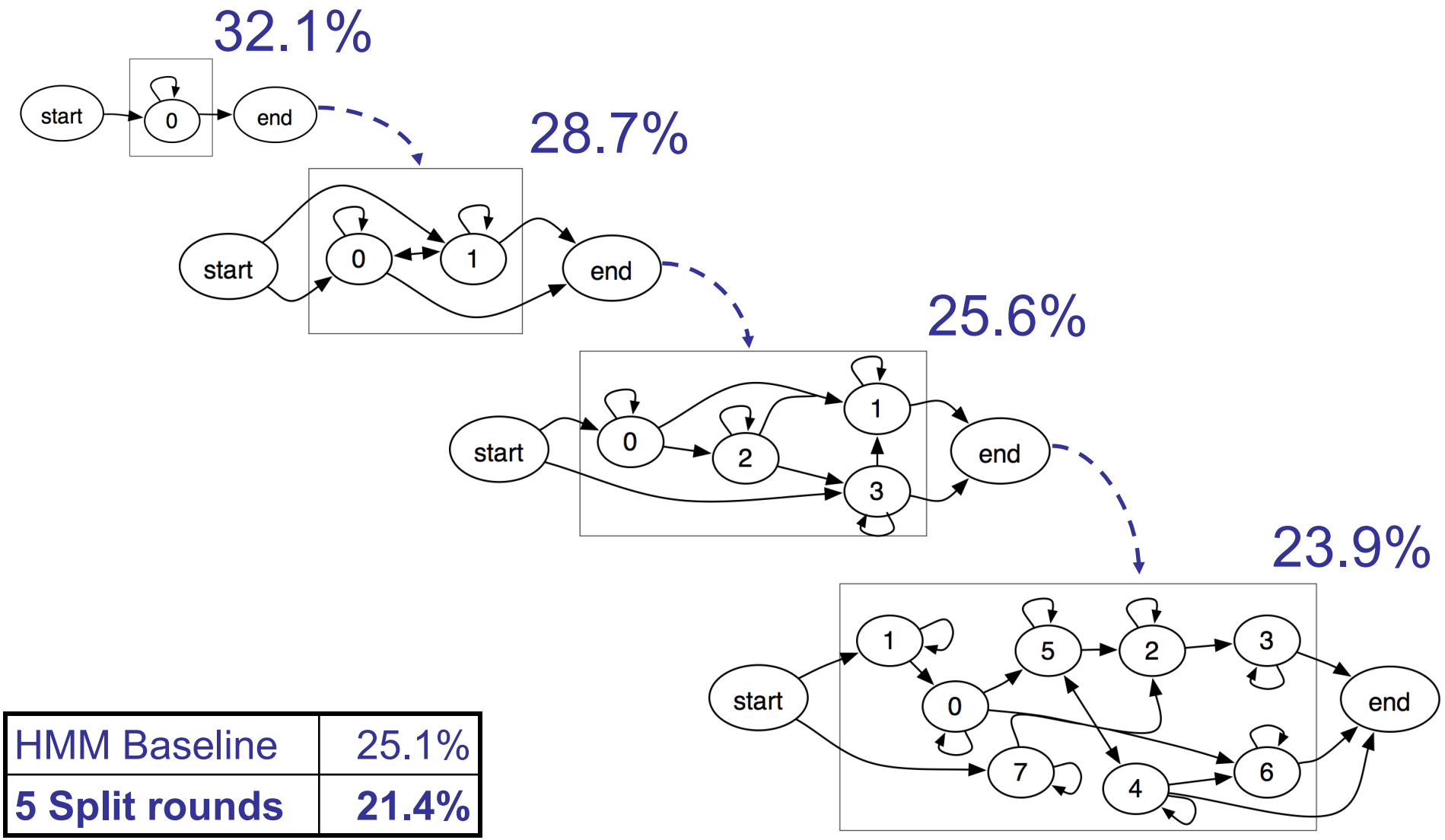
Fully Connected



Single Gaussians

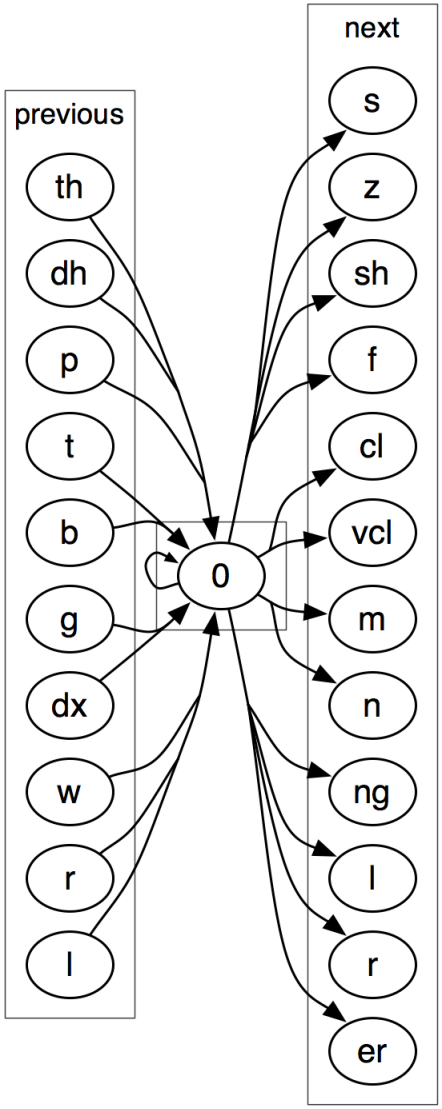
[Petrov, Pauls, and Klein, 07]

Hierarchical Split Training with EM

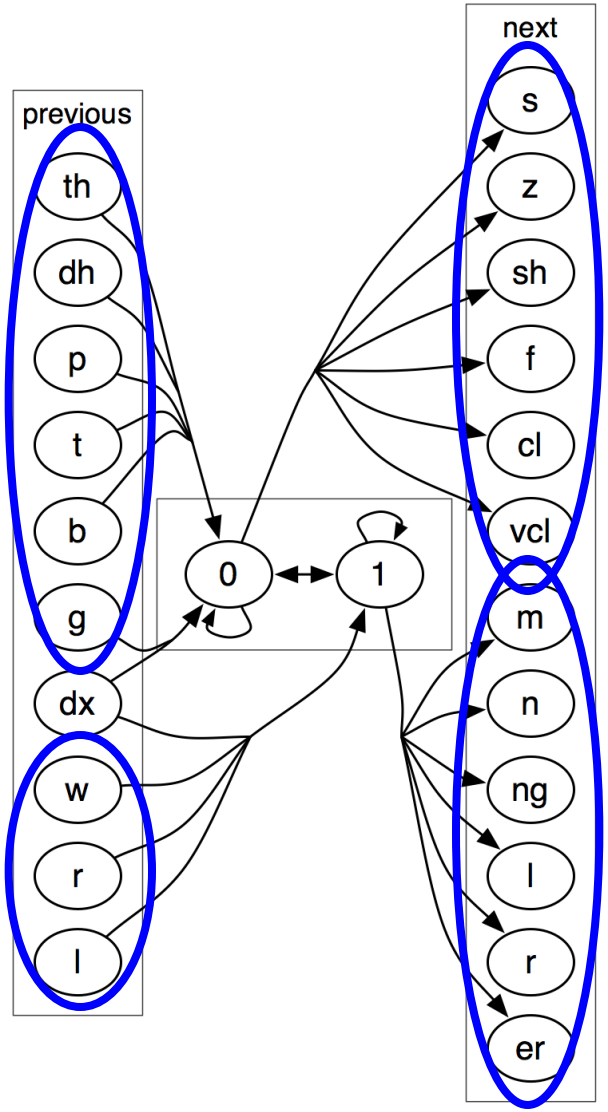


HMM Baseline	25.1%
5 Split rounds	**21.4%**

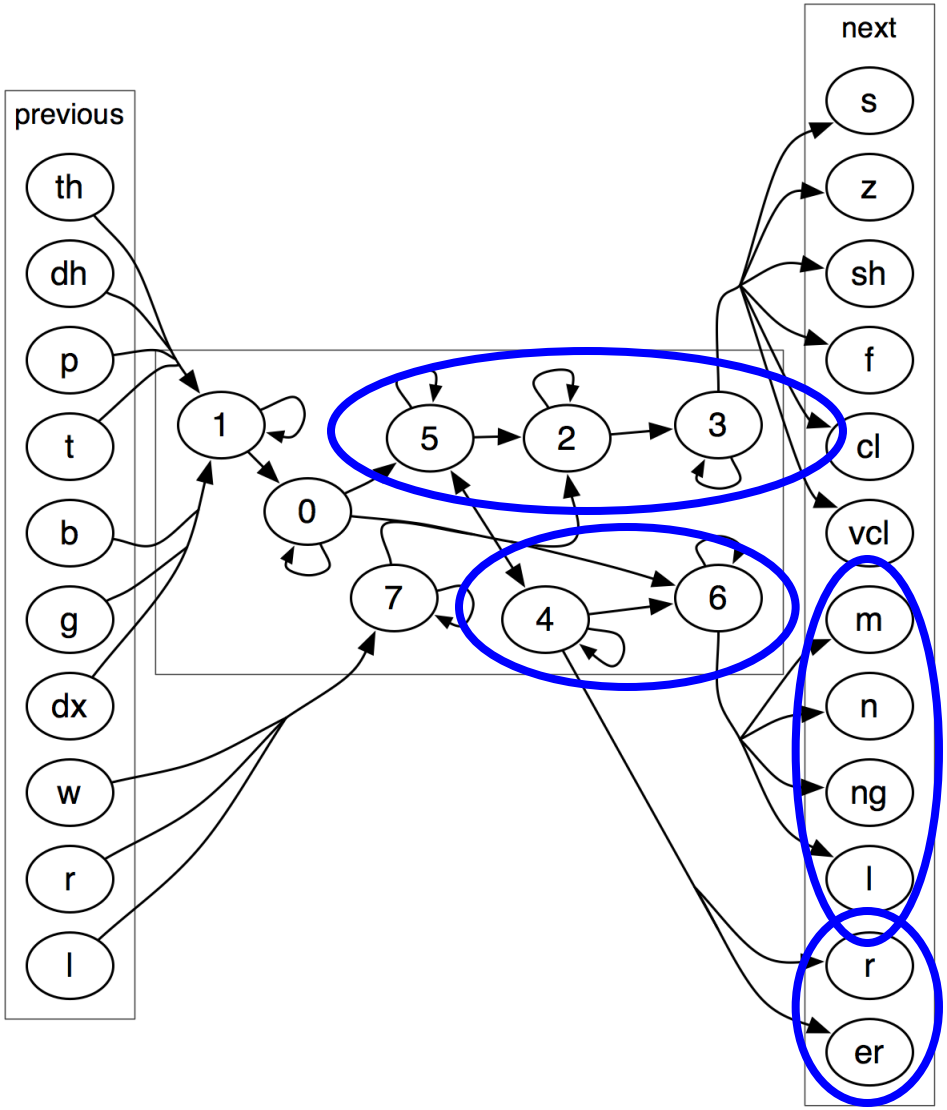
Refinement of the /ih/-phone



Refinement of the /ih/-phone



Refinement of the /ih/-phone



HMM states per phone