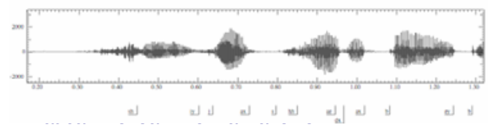
Statistical NLP
Spring 2007

Lecture 9: Acoustic Models

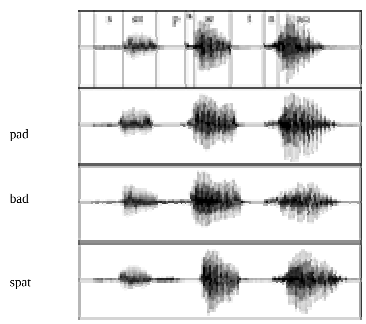
Dan Klein – UC Berkeley

She just had a baby

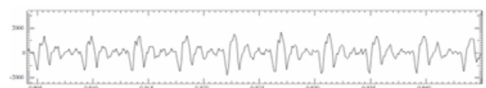


- Vowels are voiced, long, loud
- Length in time = length in space in waveform picture
- Voicing: regular peaks in amplitude
- When stops closed: no peaks: silence.
- Peaks = voicing: .46 to .58 (vowel [iy], from second .65 to .74 (vowel [ax]) and so on
- Silence of stop closure (1.06 to 1.08 for first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

Examples from Ladefoged



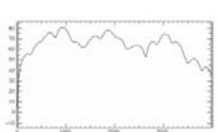
Part of [ae] waveform from "had"



- 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 Spectra

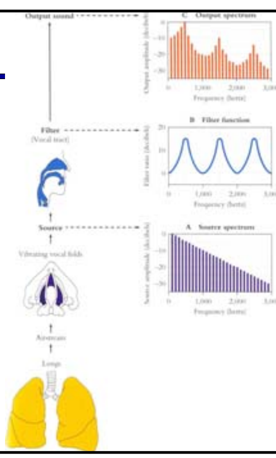
- Spectrum represents these freq components
- Computed by Fourier transform, algorithm which separates out each frequency component of wave.



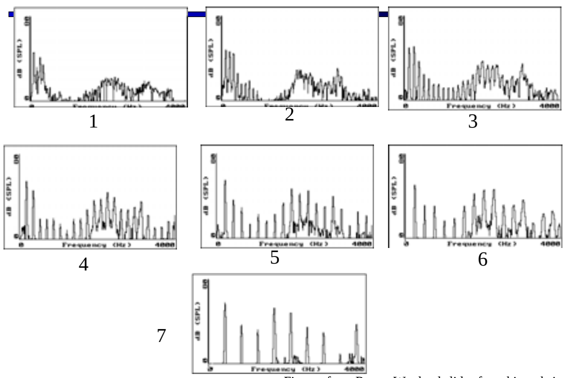
- x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

Why these Peaks?

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



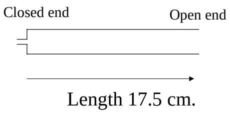
Vowel [i] sung at successively higher pitch.



Figures from Ratree Wayland slides from his website

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.

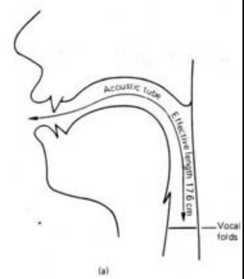
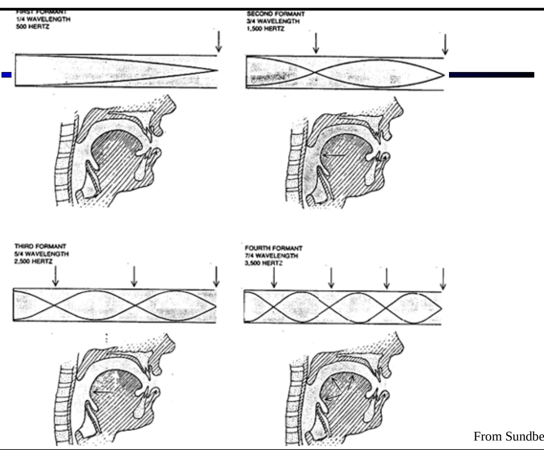


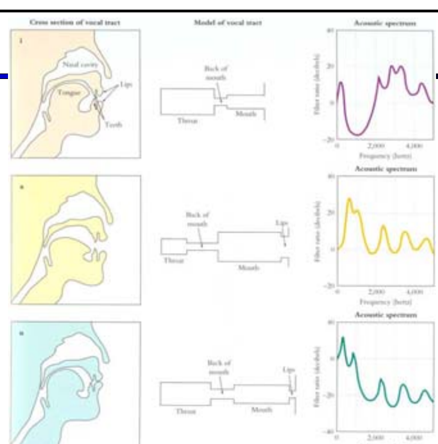
Figure from W. Barry Speech Science slides



From Sundberg

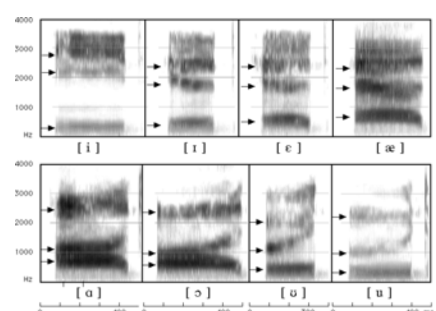
Computing the 3 Formants of Schwa

- 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's Web site

Formants in Spectrograms



American English Vowel Space

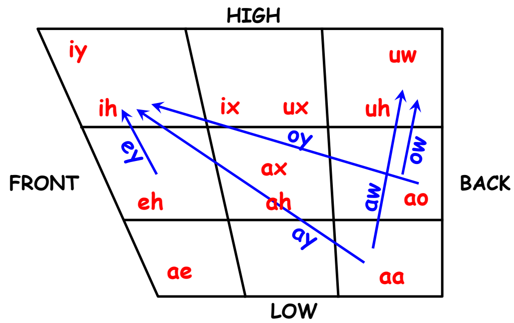
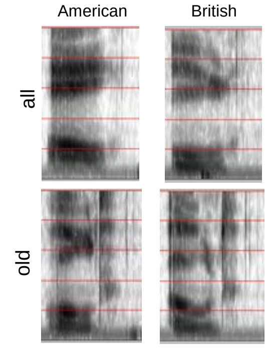


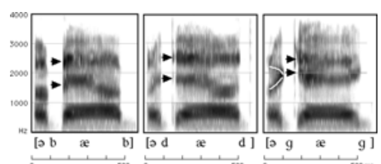
Figure from Jennifer Venditti

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 (butter: [♌✍☼❒] vs. [♌✍♦☜])
  - Phonetic
- Mismatch between training and testing dialects can cause a large increase in error rate



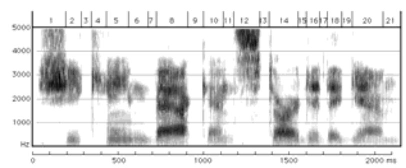
Stops in Spectrograms

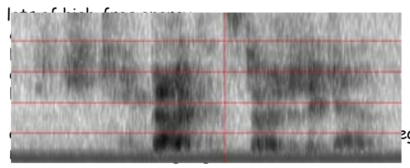


- bab: **closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab"**
- dad: **first formant increases, but F2 and F3 slight fall**
- gag: **F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials**

From Ladefoged "A Course in Phonetics"

She came back and started again



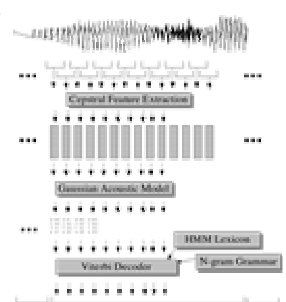


From Ladefoged "A Course in Phonetics"

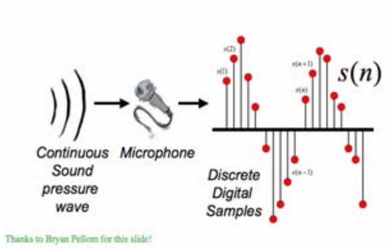
The Noisy Channel Model

- Search through space of all possible sentences.
- Pick the one that is most probable given the waveform.

Speech Recognition Architecture



Digitizing Speech



Frame Extraction

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

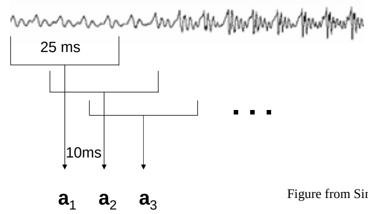


Figure from Simon Arnfield

Mel Freq. Cepstral Coefficients

- Do FFT to get spectral information
  - Like the spectrogram/spectrum we saw earlier
- Apply Mel scaling
  - Linear below 1kHz, log above, equal samples above and below 1kHz
  - Models human ear; more sensitivity in lower freqs
- Plus Discrete Cosine Transformation

Final Feature Vector

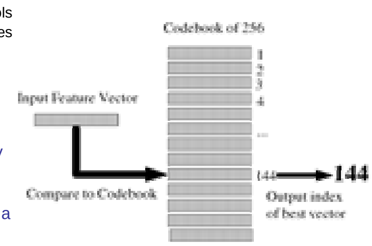
- 39 (real) features per 10 ms frame:
  - 12 MFCC features
  - 12 Delta MFCC features
  - 12 Delta-Delta MFCC features
  - 1 (log) frame energy
  - 1 Delta (log) frame energy
  - 1 Delta-Delta (log frame energy)
- So each frame is represented by a 39D vector

HMMs for Continuous Observations?

- Before: discrete, finite set of observations
- Now: spectral feature vectors are real-valued!
- Solution 1: discretization
- Solution 2: continuous emissions models
  - 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; too simple
- Useful to consider as a starting point



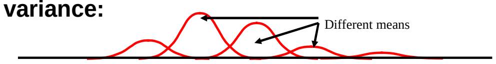
Gaussian Emissions

- VQ is insufficient for real ASR
- Instead: Assume the possible values of the observation vectors are normally distributed.
- Represent the observation likelihood function as a Gaussian with mean μ_j and variance σ_j^2

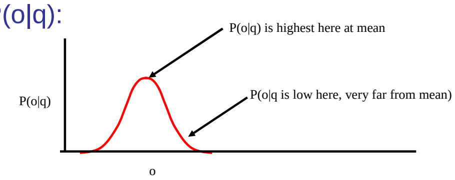
$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Gaussians for Acoustic Modeling

A Gaussian is parameterized by a mean and a variance:



- P(o|q):



Multivariate Gaussians

- Instead of a single mean μ and variance σ:

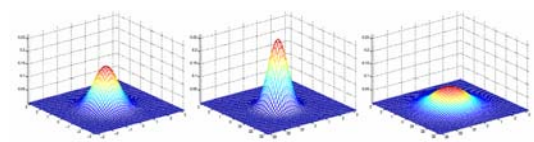
$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Vector of means μ and covariance matrix Σ

$$f(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

- Usually assume diagonal covariance
  - This isn't very true for FFT features, but is fine for MFCC features

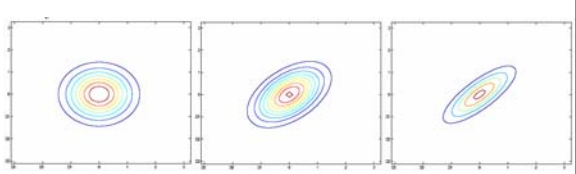
Gaussian Intuitions: Size of Σ



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

Text and figures from Andrew Ng's lecture notes for CS229

Gaussians: Off-Diagonal

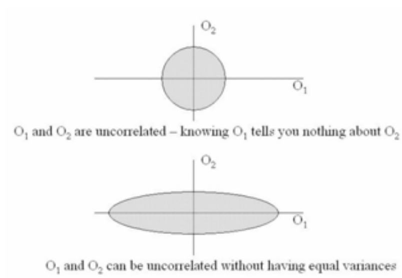


$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

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

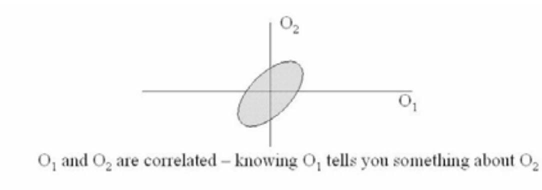
Text and figures from Andrew Ng's lecture notes for CS229

In two dimensions



From Chen, Picheny et al lecture slides

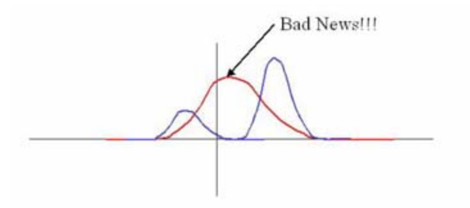
In two dimensions



From Chen, Picheny et al lecture slides

But we're not there yet

- Single Gaussian may do a bad job of modeling distribution in any dimension:



- Solution: Mixtures of Gaussians

Figure from Chen, Picheney et al slides

Mixtures of Gaussians

- M mixtures of Gaussians:

$$f(x \mid \mu_{jk}, \Sigma_{jk}) = \sum_{k=1}^{M} c_{jk} N(x, \mu_{jk}, \Sigma_{jk})$$

$$b_j(o_t) = \sum_{k=1}^{M} c_{jk} N(o_t, \mu_{jk}, \Sigma_{jk})$$

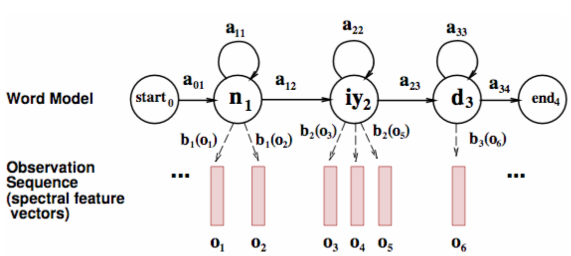
- For diagonal covariance:

$$b_j(o_t) = \sum_{k=1}^{M} \frac{c_{jk}}{2\pi^{D/2} \prod_{d=1}^{D} \sigma_{jkd}^{2}} \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \frac{(x_{jkd} - \mu_{jkd})^2}{\sigma_{jkd}^{2}}\right)$$

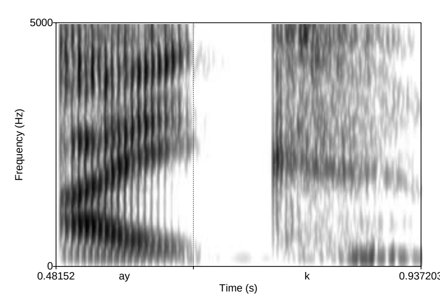
GMMs

- Summary: each state has a likelihood function parameterized by:
  - M mixture weights
  - M mean vectors of dimensionality D
  - Either
    - M covariance matrices of DxD
  - Or often
    - M diagonal covariance matrices of DxD which is equivalent to
    - M variance vectors of dimensionality D

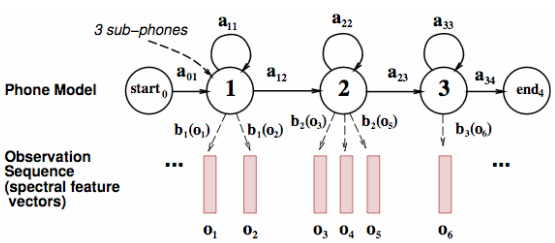
HMMs for Speech



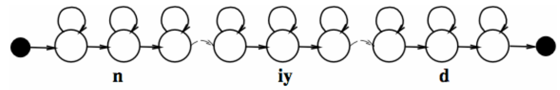
Phones Aren't Homogeneous



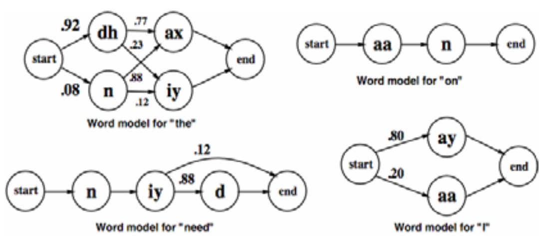
Need to Use Subphones



A Word with Subphones



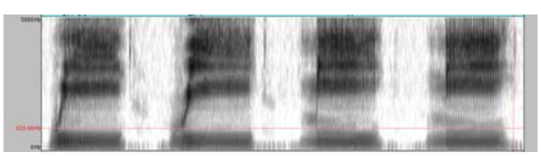
ASR Lexicon: Markov Models



Training Mixture Models

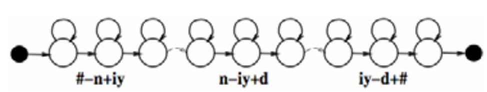
- Forced Alignment
  - Computing the "Viterbi path" over the training data is called "forced alignment"
  - We know which word string to assign to each observation sequence.
  - We just don't know the state sequence.
  - So we constrain the path to go through the correct words
  - And otherwise do normal Viterbi
- Result: state sequence!

Modeling phonetic context



W iy r iy m iy n iy

"Need" with triphone models

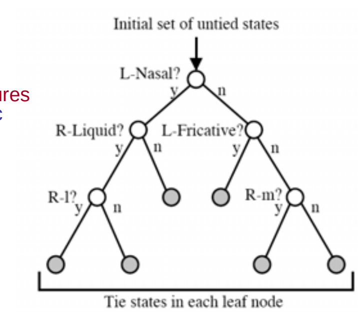


Implications of Cross-Word Triphones

- 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
  - But in training data only 22,800 triphones occur!
- Need to generalize models.

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



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