Frame Extraction

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

![Frame Extraction Diagram](https://example.com/frame-diagram.png)

Figure from Simon Arnfield

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

![Vector Quantization Diagram](https://example.com/vq-diagram.png)

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 $\mu_i$ and variance $\sigma_i^2$

$$f(x | \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$
Gaussians for Acoustic Modeling

A Gaussian is parameterized by a mean and a variance:

- \( P(o|q) \):
  - \( P(o|q) \) is highest here at mean
  - \( P(o|q) \) is low here, very far from mean

- Different means

Multivariate Gaussians

- Instead of a single mean \( \mu \) and variance \( \sigma \):
  \[
  f(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)
  \]

- Vector of means \( \mu \) and covariance matrix \( \Sigma \):
  \[
  f(x | \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right)
  \]

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

Gaussian Intuitions: Size of \( \Sigma \)

- \( \mu = [0 \ 0] \)
- \( \Sigma = I \)
- \( \Sigma = 0.6I \)
- \( \Sigma = 2I \)

- As \( \Sigma \) becomes larger, Gaussian becomes more spread out; as \( \Sigma \) becomes smaller, Gaussian more compressed

Gaussians: Off-Diagonal

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

In two dimensions

- \( O_1 \) and \( O_3 \) are correlated – knowing \( O_1 \) tells you something about \( O_3 \)
- \( O_1 \) and \( O_3 \) can be uncorrelated without having equal variance

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:

  ![Image](image.png)

- Solution: Mixtures of Gaussians

Mixtures of Gaussians

- M mixtures of Gaussians:

  \[
  f(x | \mu_j, \Sigma_j) = \sum_{k=1}^{M} c_{jk} N(x; \mu_{jk}, \Sigma_{jk})
  \]
  
  \[
  b_j(o_t) = \sum_{k=1}^{M} c_{jk} N(o_t; \mu_{jk}, \Sigma_{jk})
  \]

- For diagonal covariance:

  \[
  b_j(o_t) = \sum_{k=1}^{M} c_{jk} \frac{1}{\sqrt{2\pi\sigma_{jk}}} \exp\left(-\frac{1}{2} \frac{(x_{ot} - \mu_{jk})^2}{\sigma_{jk}^2}\right)
  \]

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

Viterbi Decoding

ASR Lexicon: Markov Models

Markov Process with Bigrams

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!
“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)

Standard subphone/mixture HMM

Our Model

<table>
<thead>
<tr>
<th>Model</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM Baseline</td>
<td>25.1%</td>
</tr>
</tbody>
</table>
Hierarchical Baum-Welch Training

Refinement of the /ih/-phone

Refinement of the /ih/-phone

Refinement of the /ih/-phone

HMM states per phone

Inference

- State sequence: \( d_1, d_2, d_3, d_4, a_e, d_5, d_6, d_7, d_8, d_9, d_{10} \)
- Phone sequence: \( d, d, d, d, a_e, a_e, a_e, d, d, d, d, d \)
- Transcription: \( d, a_e, a_e, d \)

Viterbi

Variational

???