Return to Main

Objectives

Mixture Generation:
EM Estimation
Clustering
Variance-Splitting

Temporal Modeling:
Independence
Duration
First-Order

Review:
Syllabus

On-Line Resources:
Clustering
Conditional Independence
Ten Years of HMMs

LECTURE 28: PRACTICAL ISSUES

- Objectives:
  - Mixture splitting
  - Clustering
  - Conditional independence
  - Duration modeling
  - Higher order processes

This lecture combines material from the course textbook:

X. Huang, A. Acero, and H.W. Hon, *Spoken Language Processing - A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Upper Saddle River, New Jersey, USA, ISBN: 0-13-022616-5, 2001.

and from this source:

S.Young, *et al*, *The HTK Book (v3.0)*, Cambridge University Engineering Department, September 2000.

LECTURE 28: PRACTICAL ISSUES

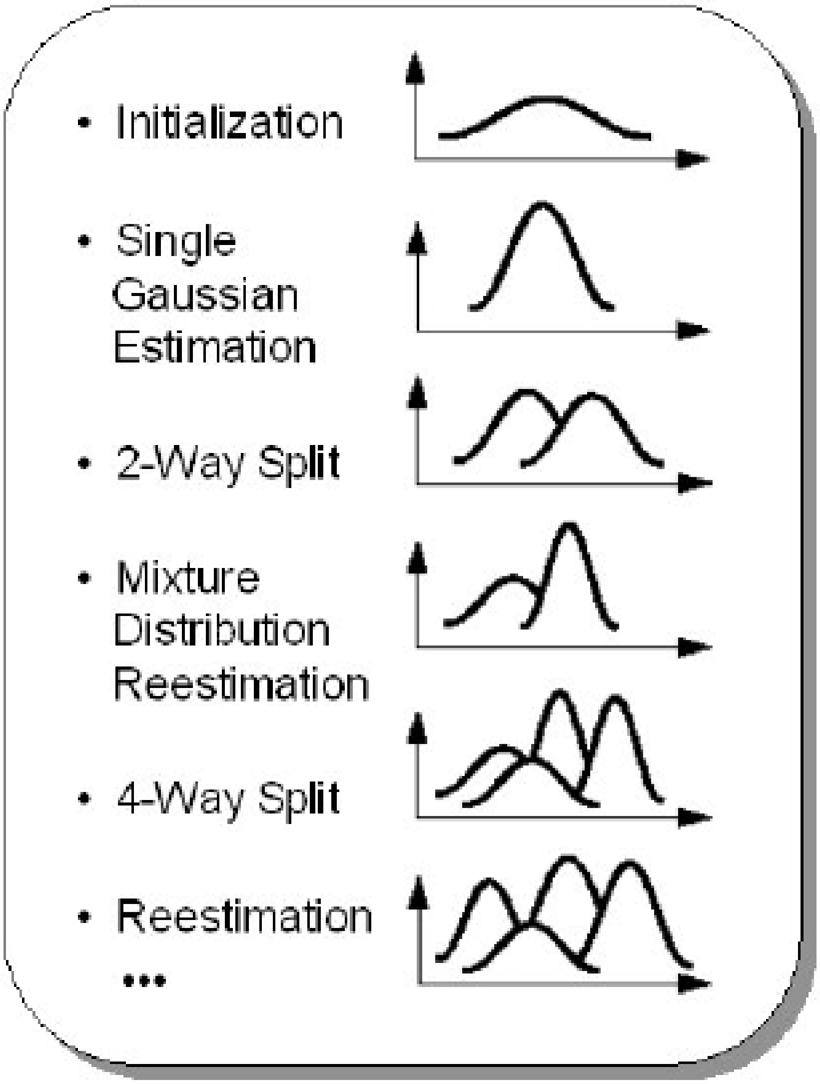
- Objectives:
  - Mixture splitting
  - Clustering
  - Conditional independence
  - Duration modeling
  - Higher order processes

This lecture combines material from the course textbook:

> X. Huang, A. Acero, and H.W. Hon, *Spoken Language Processing - A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Upper Saddle River, New Jersey, USA, ISBN: 0-13-022616-5, 2001.

and from this source:

> S.Young, *et al*, *The HTK Book (v3.0)*, Cambridge University Engineering Department, September 2000.



- Closed-loop data-driven modeling supervised only from a word-level transcription
- The expectation/maximization (EM) algorithm is used to improve our parameter estimates.
- Computationally efficient training algorithms (Forward-Backward) have been crucial.
- Batch mode parameter updates are typically preferred.
- Decision trees are used to optimize parameter-sharing, system complexity, and the use of additional linguistic knowledge.

K-MEANS CLUSTERING

Algorithm Overview:

- **Initialization**: Choose K centroids
- **Recursion**:
  - Assign all vectors to their nearest neighbor.
  - Recompute the centroids as the average of all vectors assigned to the same centroid.
- **Termination**: Check overall distortion.

For a typical implementation of K-MEANS, see our pattern recognition applet.

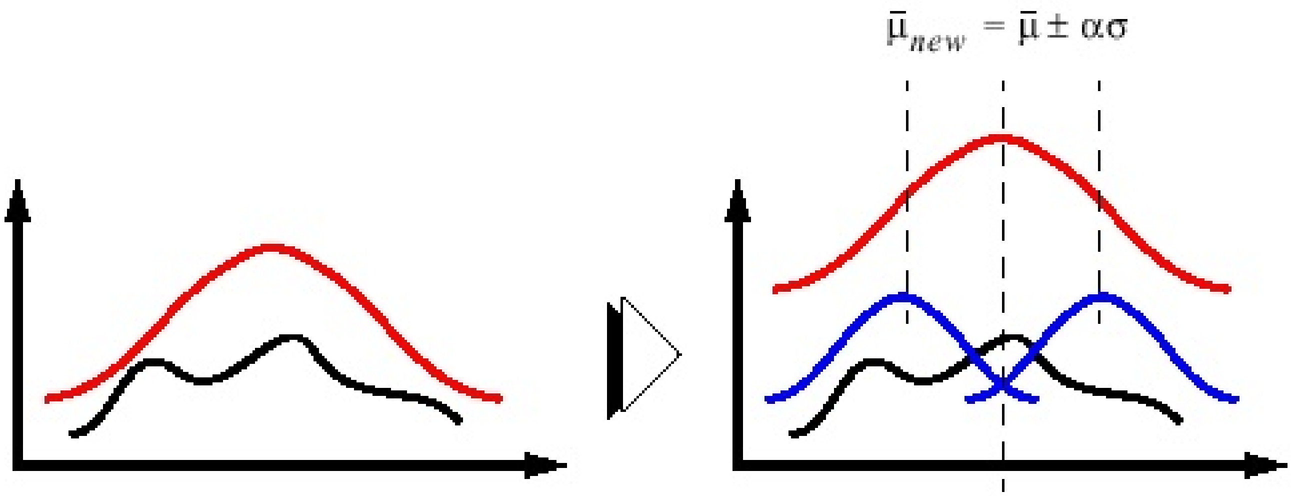
Issues:

- **Distance measure**: Euclidean? Mahalanobis?
- **Centroid computation**: Average? Median? Min-Max?
- **Splitting/Merging**: Sparsity? Separability?
- **Number of clusters**: When do we stop?

TREE-BASED CLUSTERING: VARIANCE-SPLITTING

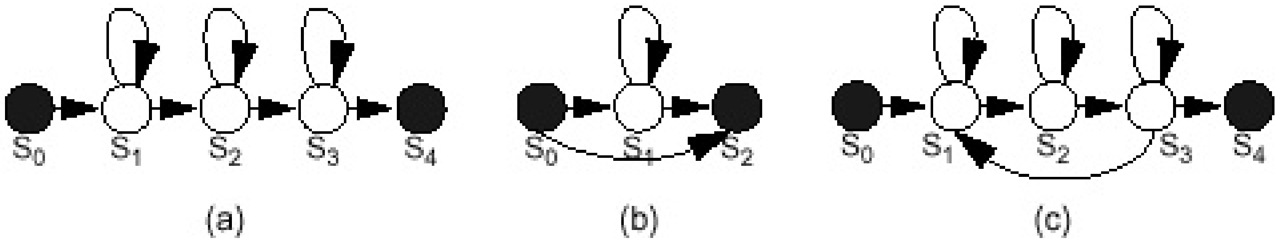
Algorithm Overview:

- Iteratively split the Gaussian with the highest mixture weight.
- Perturb the mean by a fraction of the variance:



HMM LIMITATIONS: CONDITIONAL INDEPENDENCE

Recall our basic acoustic model topology:



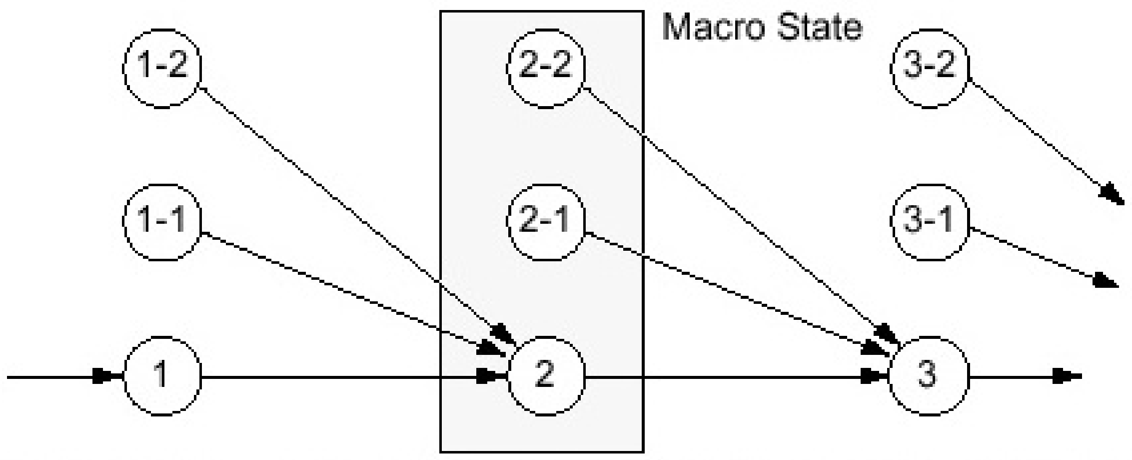
It can be argued that HMMs do not provide a realistic model for the temporal structure of speech:

- Observation probabilities for each frame (or state) are independent of previous or future frames (conditional independence). Is this a realistic model?
- The probability of staying in a state decays exponentially.

What can we do to overcome these deficiencies?

DURATION MODELING

We can explicity model duration using an alternate acoustic model topology:



We can derive suitable reestimation equations for a probability density function at each state:

Let $d_i(\tau)$ be the probability of staying in a state i for τ frames. The transition probability from state i at time t to state j at time $t+\tau$, denoted by $\gamma_{t,\tau}$, can be written as:

$$\gamma_{t,\tau}(j|i) = \frac{\alpha_t(i)a(j|i)d_i(\tau)\left(\prod_{l=1}^{\tau} b_j(y(t+l))\right)\beta_{t+\tau}(j)}{\sum_{k=1}^{N}\alpha_T(k)}$$

The probability of being in state j at time t with duration τ can be computed as:

$$\gamma_{t,\tau}(j) = \sum_{i=1}^{N}\gamma_{t,\tau}(j|i)$$

In practice, such refinements have given minimal improvements in performance.

HIGHER-ORDER MARKOV PROCESSES

Recall our first-order Markov process:

$$P[q_t = j | (q_{t-1} = i, q_{t-2} = k, \ldots)] = P[q_t = j | q_{t-1} = i]$$

We considered only those processes for which the right-hand side is independent of time:

$$a_{ij} = P[q_t = j | q_{t-1} = i] \qquad 1 \le i, j \le N$$

We can extend this model to account for previous transitions:

$$a_{ijk} = P[q_t = k | (q_{t-1} = i, q_{t-2} = j)] \qquad 1 \le i, j, k \le N$$

We now have a second-order Markov process. We can derive suitable maximum likelihood reestimation equations:

$$\gamma_t(k|(i, j)) = \frac{\alpha_t(i, j) a(k|(i, j)) b_k(y_t) \beta_{t+1}(i, j)}{\sum_{l=1}^{N} \alpha(y_1^T, l)}$$

However, in practice, the benefits of this model have not offset the significant increase in computational costs.