**EECS225d Audio Signal Processing in Humans and Machines** **Spring 1995**

# Lecture 33: April 12, 1995

*Lecturer: Professor Nelson Morgan* *Scribe: Kristine Ma*

## 33.1 Overview

Some of the major recognition approaches are:

- Stored Templates, DTW. In this approach, a metric space is defined. Input patterns are compared with the stored patterns by computing their relative distances. In dynamic time warping, the input is also stretched/warped nonlinearly in time to find matches with the stored pattern.
- Statistical Models, Continuous HMM. In statistical modeling, the cost is interpreted in terms of probability rather than distance. Furthermore, powerful statistical learning tools are available for this approach. Using the Viterbi criterion, this approach is analogous to DTW with shorter "templates".
- Statistical Models, Discrete HMM. In discrete HMMs, the acoustic vector sequence is quantized according to a predetermined codebook, and as in continuous HMMs, probabilities are used.
- Neural Network. This can be said to be a special case of statistical modeling, but not all NN's are build with a statistical model in mind.
- Statistical Models + Neural Network.

DTW was discussed in a previous lecture. Today we are going to cover hidden Markov models. Next lecture is going to be on neural networks.

## 33.2 Statistical Speech Recognition

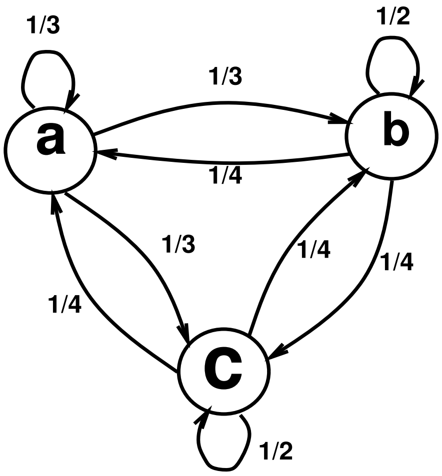
In statistical speech recognition, speech is modeled by a probabilistic sequence of states. The hidden Markov model (HMM) is the most common form of statistical model for speech or other segmented processes. Given the incoming data, the statistical approach finds the most probable model sequence that produces it. Viterbi search (dynamic programming) is usually used to find the best state sequence through each model.

## 33.3 Markov Model

A Markov model is a finite state automaton with stochastic transitions that satisfies the Markov condition.

Consider a three-state Markov model, where each state corresponds to one day's weather (assuming state residency is per day):

| state | | output |
|---|---|---|
| a | $\longleftrightarrow$ | sunny |
| b | $\longleftrightarrow$ | cloudy |
| c | $\longleftrightarrow$ | rainy |




This model has a stochastic transition matrix,

$$A = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

.

**Example 1** The sequence "a b c" occurs with a probability of $\frac{1}{3} \times \frac{1}{4} = \frac{1}{12}$.
**Example 2** Compute the probability of see the sequence "a X c", where X could be anything:

$$\begin{array}{lll} p(a_1 \rightarrow b_2, b_2 \rightarrow c_3) & = \frac{1}{3} \times \frac{1}{4} & = \frac{1}{12} \\ p(a_1 \rightarrow a_2, a_2 \rightarrow c_3) & = \frac{1}{3} \times \frac{1}{4} & = \frac{1}{12} \\ p(a_1 \rightarrow c_2, c_2 \rightarrow c_3) & = \frac{1}{4} \times \frac{1}{2} & = \frac{1}{8} \\ \hline p(a_1 \rightarrow x, x \rightarrow c_3) & & = \frac{7}{24} \end{array}$$

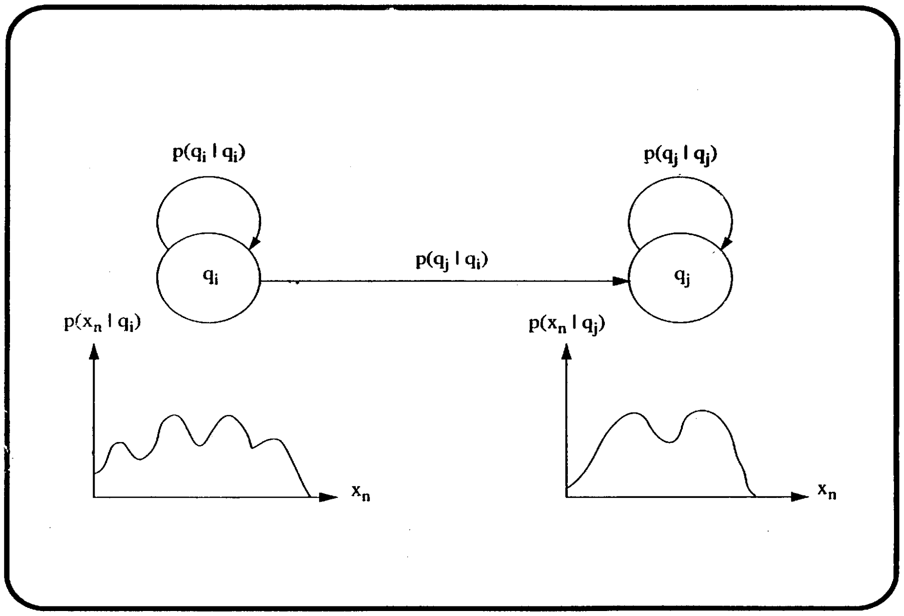
In general, the probability of a sequence is computed by summing over all possible paths, where each path is a product of transition (we will see that for HMMs, emission probabilities will also be required) probabilities.

## 33.4 Hidden Markov Model

Each output of a Markov model corresponds to a deterministic event, whereas each output of a hidden Markov model (HMM) corresponds to a probabilistic density function; thus the actual state sequence of a HMM is "hidden". As an example, imagine there are three urns hidden behind a veil. Each urn contains a certain probability distribution of white, red, and blue balls. At each instant of time, an user randomly select an urn and pick a ball from it with replacement. What the audience observed is the color of the balls, but not the urn (corresponding to the states of the HMM) from which they are selected from.

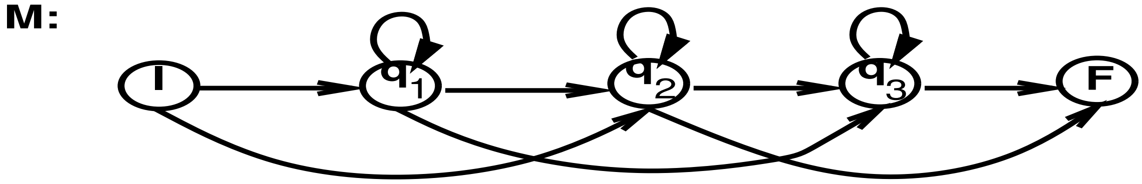
In speech recognition, the observables are the speech feature vectors (color of the balls). What we want to find out is the sequence of linguistic classes (the order of the urn from which the balls were selected).

A two-state left-to-right HMM is illustrated in the figure below. Each state $q_i$ is associated with a distribution function (emission pdf) $p(x_n|q_i)$, a self-transition probability $p(q_i|q_i)$, and a transition probability of $p(q_j|q_i)$ going to state $q_j$. The density function associated with each state is usually assumed to be Gaussian or a mixture of Gaussians. In this case, for simplicity the feature vector is a single component, such as the centroid of the speech spectrum for that frame. Again, the model is called "hidden" because the state sequence is not observed; rather the emitted variables are observed.




## 33.5 HMM in Speech Recognition

Here is another example of a HMM. This five-state model consists of two non-emitting initial(I) and final(F) states, and three emitting states. Each state occupancy in the model represents a quantization in time. It usually represents a phonemic or sub-phonemic unit, where each phoneme duration can vary according to the self-loop transition probabilities. In speech a sequence of phonemes makes up a word, and a sequence of words makes up a sentence.



### 33.5.1 Definition of Parameters

$Q = q_1, ..., q_k, ..., q_L$= set of possible HMM states
$I$ = non-emitting initial state
$F$ = non-emitting final state
$q^n$ = specific state $\in$ Q visited at time $n$
$Q_m^n = q^m, q^{m+1}, ..., q^n$
$q_k^n$ = states $q_k$ visited at time $n$
$X = x_1, ..., x_n, ..., x_N$ = sequence of acoustic vectors
$X_m^n = x_m, x_{m+1}, ..., x_n$
$M$ = the model = finite state stochastic automata.

### 33.5.2 Bayes Rule

For speech recognition, Bayes rule states that

$$P(M|X) = \frac{p(X|M)P(M)}{p(X)},$$

where $P(M)$ is the prior and is typically given by the language model, $p(X)$ (the probability of observing a particular feature vector) is constant during recognition, and $p(X|M)$ is the likelihood of the acoustic model.

### 33.5.3 Problems

#### 33.5.3.1 Definition of parameters associated with each state $q_i$

- Transition probabilities $p_{ij}$, where $\sum_j p_{ij} = 1$.
- Emission probabilities $p(x_n|q_i)$.
  - If we assume the emission probability can be modeled by a continuous Gaussian pdf, then

    $$p(x_n|q_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} exp[-\frac{(x_n - \mu_i)^t \Sigma_i^{-1}(x_n - \mu_i)}{2}],$$

    where $x_n$ is a $d$-component column vector, $\mu_i$ is the $d$-component mean vector, $\Sigma_i$ is the $d$-by-$d$ covariance matrix, $(x_n - \mu_i)^t$ is the transpose of $x_n - \mu_i$, $\Sigma^{-1}$ is the inverse of $\Sigma$, and $|\Sigma|$ is the determinant of $\Sigma$.
  - Modeling the emission probability by a discrete pdf. The first step is to obtain a set of prototype vectors $c_1, c_2, ..., c_k$ by clustering (K-means). Then $p(x_n|q_i) = p(c_k|q_i)$ where $c_k$ is such that $d(x_n, c_k) < d(x_n, c_j), \forall j \neq k$. This method does not assume that the pdf is of Gaussian distribution, but one has to determine the number of prototype vectors and the distance measure for quantization.

#### 33.5.3.2 Estimation of $P(X|M)$

**Maximum Likelihood**

The most straightforward way of calculating the probability of observing the acoustic vector $X = x_1, ..., x_n, ..., x_N$ given the model $M$ is by enumerating every possible state sequence of length $N$, $Q = q^1, ..., q^n, ..., q^N$, and then summing the probability of $X$ being generated by each of the possible state sequence $\mathcal{Q}$:

$$P(X|M) = \sum_{all\ Q\ of\ length\ N} P(X_1^N|Q, M)P(Q|M)$$

Though straight forward, this procedure requires on the order of $2N \times L^N$ calculations, where $L$ is the number of emitting states of the model (there are $L^N$ possible state sequences and each state sequence requires about 2N calculations). Fortunately, a more efficient procedure exists and it is called the "Forward-Backward" algorithm. This procedure basically divides the joint probability into two sections:

- The forward part defines the probability of the joint event that sequence $X_1^n$ (until time $n$) is observed, and the state at time $n$ is $q_l$, given the model $M$,

$$\alpha_n(l) = P(X_1^n, q_l^n|M).$$

- The backward part defines the probability of the partial observation sequence from $n+1$ to the end, given state $q_l$ at time $n$, the pass observation history, and the model $M$,

$$\beta_n(l) = P(X_{n+1}^N|q_l^n, X_1^n, M).$$

Hence,

$$P(X|M) = \sum_{l=1}^{L} P(q_l^N, X_1^N|M) = \sum_{l=1}^{L} \underbrace{P(q_l^n, X_1^n|M)}_{\substack{\text{"forward"} \\ \alpha}} \underbrace{P(X_{n+1}^N|q_l^n, X_1^n, M)}_{\substack{\text{"backward"} \\ \beta}}$$

Forward recurrence:

$$P(q_l^n, X_1^n|M) = \sum_{k=1}^{L} P(q_k^{n-1}, X_1^{n-1}|M)P(q_l^n, X_n|q_k^{n-1}, X_1^{n-1}, M).$$

Also, since $\beta_N(l) \;=\; 1\ or\ 0,\ \forall\ l\ \in\ [1, L]$,

$$P(X|M) = \sum_{\forall l \in F} \alpha_N(l),$$

where $F$ is all legal final states.

In practice, we generally make the following assumptions to simplify computation.

- State chain is assumed to be first order Markov. This means that the probability that the Markov chain is in state $q_l$ at time $n$ depends only on the state of the Markov chain at time $n-1$, and is conditionally independent of the past.

$$P(q_l^n|q_k^{n-1}, X_1^{n-1}, M) \rightarrow P(q_l^n|q_k^{n-1}, M)$$

- Observations assumed to be independent of past features and states. This means that the probability that a particular acoustic vector will be emitted at time $n$ depends only on the transition taken at that time, and is conditionally independent of the past.

$$P(x_n|q_l^n, q_k^{n-1}, X_1^{n-1}, M) \rightarrow P(x_n|q_l^n, q_k^{n-1}, M)$$
$$P(x_n|q_l^n, q_k^{n-1}, M) \rightarrow P(x_n|q_l^n, M)$$

With these simplifications,

$$P(q_l^n, X_n | q_k^{n-1}, X_1^{n-1}, M) = P(X_n | q_l^n, q_k^{n-1}, X_1^{n-1}, M) P(q_l^n | q_k^{n-1}, X_1^{n-1}, M)$$

becomes

$$P(q_l^n, X_n | q_k^{n-1}, X_1^{n-1}, M) = P(X_n | q_l^n, M) P(q_l^n | q_k^{n-1}, M).$$

Thus the forward recurrence becomes,

$$P(q_l^n, X_1^n | M) = \left[ \sum_{k=1}^{L} P(q_k^{n-1}, X_1^{n-1} | M) P(q_l^n | q_k^{n-1}, M) \right] P(X_n | q_l^n, M).$$

**Viterbi criterion - estimation of best path**

In practice, the statistical computation is usually further simplified to perform recognition based on the best path, rather on the sum of all possible state sequences through a model. This simply involves replacing all the summations with the max function. Hence,

$$\overline{P}(X|M) = \max_{Q_1^N} P(Q_1^N, X_1^N | M) = \max_{Q_1^N} P(q_l^n, X_1^n | M) P(X_{n+1}^N | q_k^n, X_1^n, M),$$

and the forward recurrence is approximated as

$$\overline{P}(q_l^n, X_1^n | M) = \max_k P(q_k^{n-1}, X_1^{n-1} | M) P(q_l^n, X_n | q_k^{n-1}, X_1^{n-1}, M).$$

With the independence and the first order assumptions, it further reduced to

$$\overline{P}(q_l^n, X_1^n | M) = \max_k \left[ P(q_k^{n-1}, X_1^{n-1} | M) P(q_l^n | q_k^{n-1}, M) \right] P(X_n | q_l^n, M).$$

#### 33.5.3.3 Training of the parameters

If $X_j$ is one of the $J$ utterances associated with HMM $M_j$, then the parameters are estimated by maximizing $R$ over the entire parameter space $\Theta$, where for the maximum likelihood criterion, $R$ is

$$R = \prod_{j=1}^{J} P(X_j | M_j, \Theta_j),$$

and for the Viterbi criterion $R$ is

$$R = \prod_{j=1}^{J} \overline{P}(X_j | M_j, \Theta_j).$$

and $\Theta_j$ is the parameter set of $M_j$. There's no exact solution for this problem. In practice, training is done using an iterative optimization approach. The Baum-Welch re-estimation (Forward-Backward) is outlined as follow:

1. Assume initial parameters (e.g. transition probabilities and vector quantization emission probabilities or Gaussian parameters)

2. Compute expectations of these parameters using forward and backward recursions (see, e.g., Lee). Here is an example of re-estimating a conditional mean (mean vector — class i):

$$\hat{\mu_i} = \sum_{n=1}^{N} \frac{P(state\ i\ at\ time\ n)}{prob(state\ i)} x_n = \frac{\sum_{n=1}^{N} P(q_i^n|X_1^N)}{P(q_i|X_1^N)} x_n = \frac{\sum_{n=1}^{N} P(q_i^n|X_1^N) x_n}{\sum_{n=1}^{N} P(q_i^n|X_1^N)}$$

$$= \frac{\sum_{n=1}^{N} P(q_i^n, X_1^N) x_n}{\sum_{n=1}^{N} P(q_i^n, X_1^N)} = \frac{\sum_{n=1}^{N} \alpha_n(i)\beta_n(i) x_n}{\sum_{n=1}^{N} \alpha_n(i)\beta_n(i)}$$

   Note: this is just a ratio of summing over the $\alpha$ and $\beta$ coefficients.

3. Set parameters to these expected values.
4. Check for convergence; return to step 2 if not converged.

Here are the basic steps of the Viterbi training procedure:

1. initial segmentation
2. estimate HMM parameters
3. obtain new segmentation (dynamic programming)
4. update HMM parameters
5. go back to step 3 until convergence

#### 33.5.3.4 Recognition

Given a sentence $X$, find the best model sequence $M$ such that $P(X|M)$ is maximum (here we ignored the language model, which is incorporated as in 33.5.2).

**Maximum Likelihood Criterion**

- For isolated words:

$$X \in M_k\ if\ P(X|M_k) > P(X|M_j),\ \forall\ j \neq k$$

- For connected speech:
  - all paths are considered $\Rightarrow$ no segmentation
  - gives the optimal sequence of models by computing $P(X|M)$ $\forall$ possible model sequences $M$, but this is impossible from the CPU point of view
  - only 1 possible algorithm known:
    stack decoding, but still very heavy in spite of important pruning ( and thus risk of losing the optimal solution)

**Viterbi Criterion**

Use dynamic programming for pattern matching to find the best state sequence, where

$$Dist_{i,j} = -ln[P(x_i|q_j)]$$

$$Tcost_{jk} = -ln[P_{jk}].$$

## 33.6 Conclusion

In this lecture, we have presented the concept behind hidden Markov models and how they are being used for speech recognition. To simplify computation, there are several assumptions made in the use of HMMs that do not correspond to a good model of speech.

- State chain is assumed to be first order Markov. This means that the probability that the Markov chain is in state $q_l$ at time $n$ depends only on the state of the Markov chain at time $n-1$, and is conditionally independent of the past.
- Observations assumed to be independent of past features and states. This means that the probability that a particular acoustic vector will be emitted at time $n$ depends only on the transition taken at that time, and is conditionally independent of the past.
- Recognition is usually based on the best path (Viterbi), and not on all possible state sequences (Maximum Likelihood) through a model.
- Strong distribution assumptions. Discrete HMM often assumes independence of features, distribution implied by VQ distance measure. Continuous HMM assumes multi Gaussians distribution and speech being representable by a succession of steady-state segments.
- All parameters are not optimized simultaneously, rather they are separately estimated from language model and acoustic model