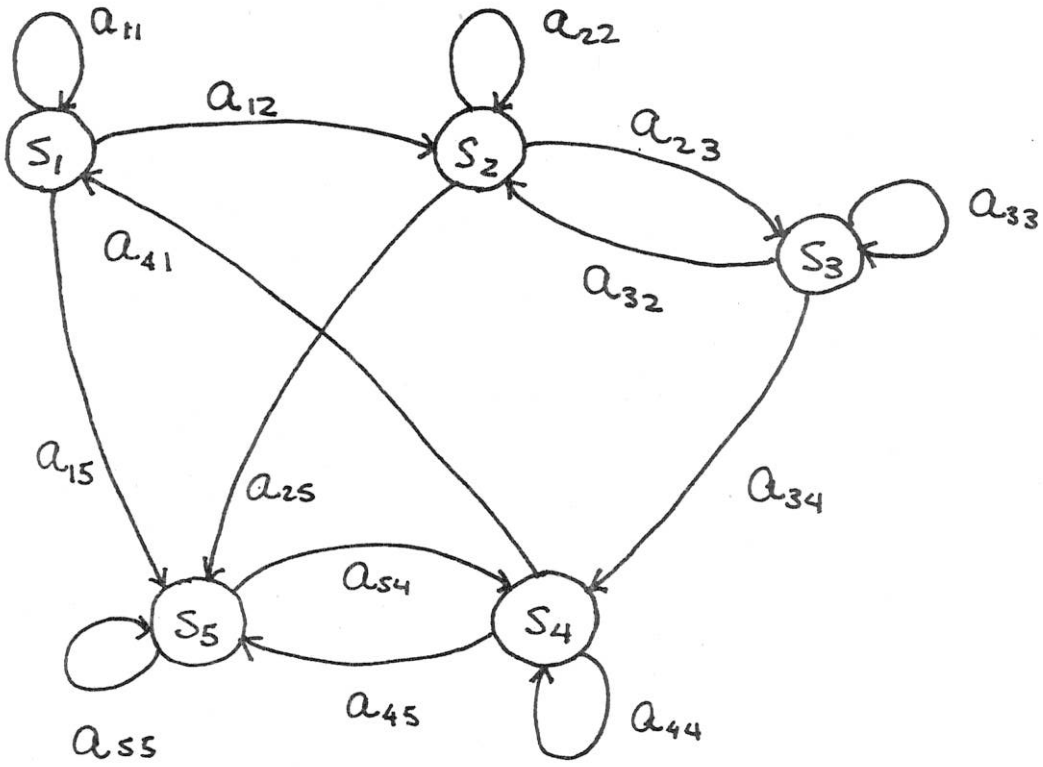
Introduction To Hidden Markov Models

Yoshiko Ito
Dragon Systems

Review of Markov Chains



- States - $S_1, S_2, \ldots$
- State transitions and associated transition probabilities a_{ij}

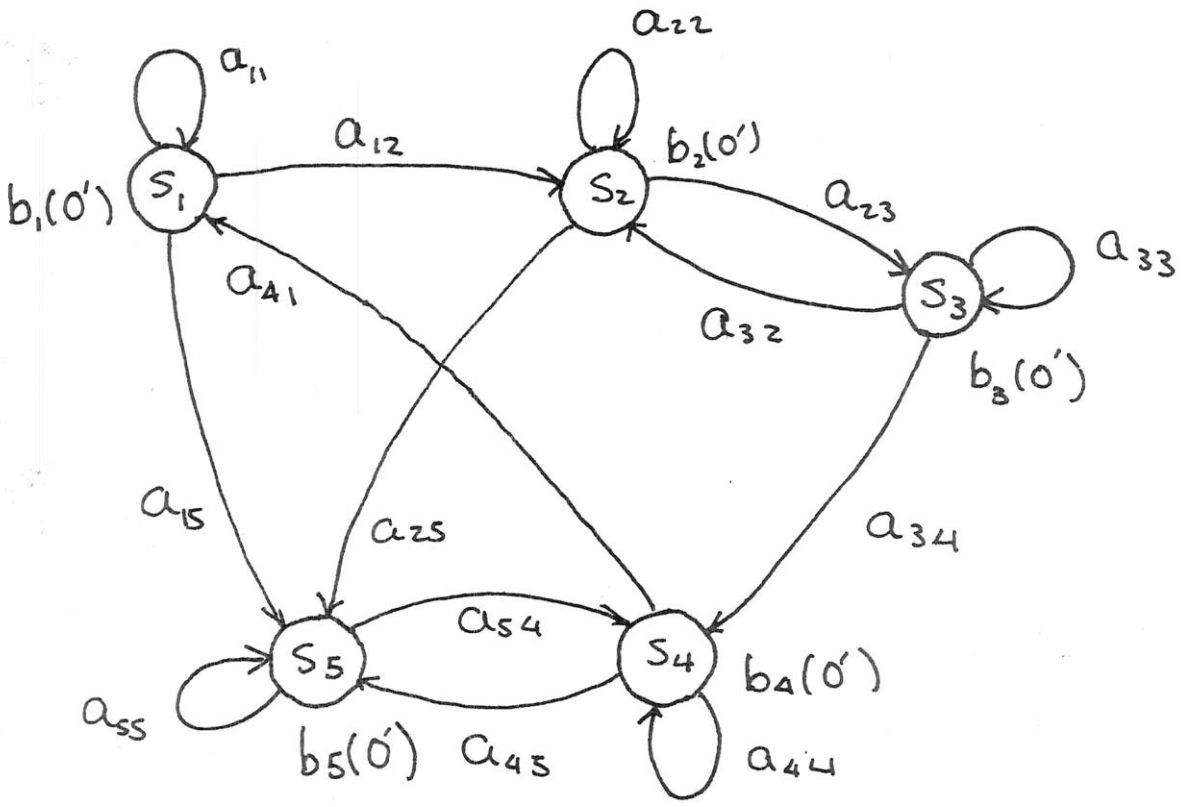
Let q_t = state at time t

$$a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)$$

Markov property:

$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, \ldots) = P(q_t = S_j \mid q_{t-1} = S_i)$$

Hidden Markov Model (HMM)



- States
- State transitions & transition probabilities
- → Output (observation) & output probabilities

$$b_i(o_t) = P(\text{observing } o_t \text{ at time } t \mid q_t = S_i)$$

- States are "hidden", i.e., they are only indirectly observable through observation sequences.
- $b_i(O_t)$ depends only on q_t and O_t

HMM Example ("Balls and Urns")

Each urn contains colored balls

Each ball is colored RED, BLUE, GREEN, or YELLOW.

URN1 URN2 URN3

	URN1	URN2	URN3
$P(R \mid URN)$	$b_1(R)=0.7$	$b_2(R)=0.2$	$b_3(R)=0.1$
	$b_1(B)=0.2$	$b_2(B)=0.6$	$b_3(B)=0.3$
	$b_1(G)=0.05$	$b_2(G)=0.1$	$b_3(G)=0.1$
	$b_1(Y)=0.05$	$b_2(Y)=0.1$	$b_3(Y)=0.5$

- Pick an urn according to a Markov process

$$a_{ij} = P(\text{picking urn } j \mid \text{last urn picked} = i)$$

- Pick a ball at random from the selected urn.
- The observer only sees the color of the selected ball.

State = urn, Observation = color of the selected ball

Three HMM Problems

1. Given an HMM and an observation sequence, find the probability that the observed sequence was generated by the model.

   Application: Recognition / Classification

2. Given an HMM and an observation sequence, find the best single state sequence that accounts for the observation.

   Application: Recognition, Segmentation

3. Given an HMM and an observation sequence, choose the model parameters so that the probability of observation is maximized.

   Application: Training

1. Given an HMM & an observation sequence find the probability that the observed sequence was generated by the model.

Let

$O_1, O_2, \cdots O_T$ = observation sequence

q_t = state at time t ; $q_t \in \{S_1, S_2 \cdots S_N\}$

$X_{q_1}, X_{q_2}, \cdots X_{q_T}$ = hidden state sequence

π_i = probability of being in state i at $t=1$

IF we knew the state sequence $X_{q_1} \cdots X_{q_T}$ then

$$P(O_1, O_2 \cdots O_T \mid X_{q_1}, \cdots X_{q_T}) = b_{q_1}(O_1) \cdots b_{q_T}(O_T)$$

But we don't know the state sequence, so add probabilities over all possible state sequences of length T.

$$P(O_1, \cdots O_T) = \overbrace{\sum_{q_1=1}^{N} \cdots \sum_{q_T=1}^{N}}^{\text{over all state sequences}} \underbrace{P(O_1 \cdots O_T \mid X_{q_1} \cdots X_{q_T})}_{\substack{\text{Prob. of observation} \\ \text{for a particular} \\ \text{State sequence}}} \overbrace{P(X_{q_1} \cdots X_{q_T})}^{\substack{\text{Probability} \\ \text{of state sequence}}}$$

Notation:

$q_t = i$ means $q_t = S_i$

Evaluate

$$P(O_1 \cdots O_T) = \sum_{q_1=1}^{N} \cdots \sum_{q_T=1}^{N} P(O_1 \cdots O_T | X_{q_1} \cdots X_{q_T}) P(X_{q_1} \cdots X_{q_T})$$

$$= \sum_{q_1=1}^{N} \cdots \sum_{q_T=1}^{N} b_{q_1}(O_1) \cdots b_{q_T}(O_T) \pi_{q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T}$$

How long would it take to do this computation?

N^T state sequences $\rightarrow$ $N^T - 1$ additions

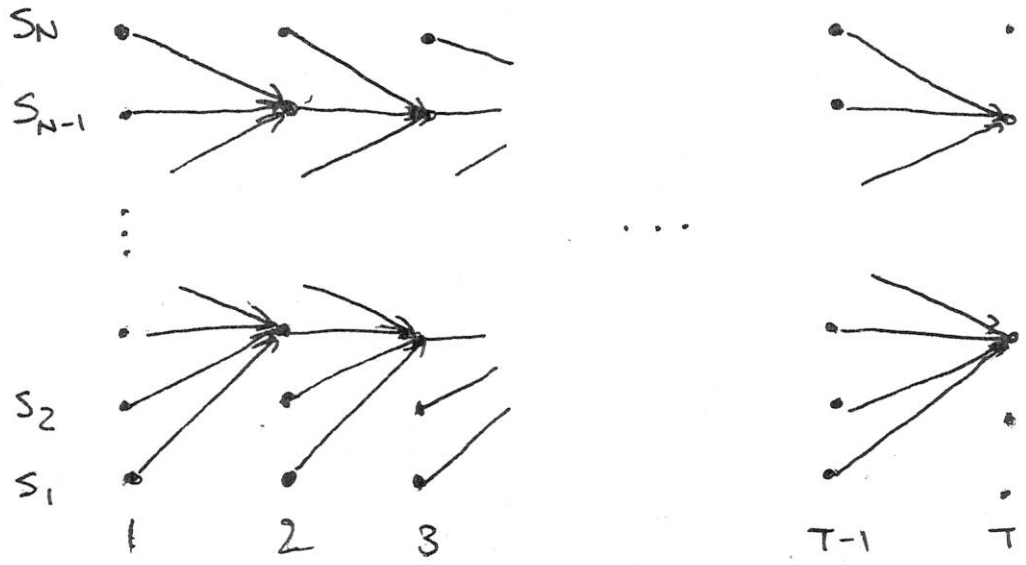
$2T-1$ multiplications / state sequence $\rightarrow$ $N^T(2T-1)$ multiplications

$\sim 2N^T T$ operations

If $N=5$, $T=100$, $2N^T T = 2 \cdot 5^{100} \cdot 100 \sim 10^{72}$

There are $\sim 3.2 \times 10^7$ seconds / year

But there is a better way!



Define forward variable $\alpha_t(i)$:

$$\alpha_t(i) = P(O_1, \cdots O_t, q_t = S_i)$$

= Prob. of being in state i at time t, having observed $O_1, \cdots O_t$

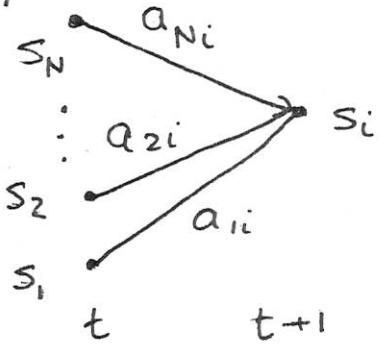
Initialization:

$$\alpha_1(i) = P(O_1, q_1 = S_i) = \underbrace{P(q_1 = S_i)}_{\pi_i} \underbrace{P(O_1 | q_1 = S_i)}_{b_i(O_1)} = \pi_i b_i(O_1)$$

Induction:

$$\alpha_{t+1}(i) = P(O_1 \cdots O_{t+1}, q_{t+1} = S_i) = P(O_1 \cdots O_t, q_{t+1} = S_i) b_i(O_{t+1})$$

$$= \left[\sum_{k=1}^{N} \alpha_t(k) a_{ki} \right] \cdot b_i(O_{t+1})$$



Termination:

$$P(O_1, \cdots O_T) = \sum_{i=1}^{N} \alpha_T(i)$$

Computation needed $O(N^2 T)$ compare with $N^T T$

Can also evaluate prob. of observation "backwards"

Define backward variable $\beta_t(i)$

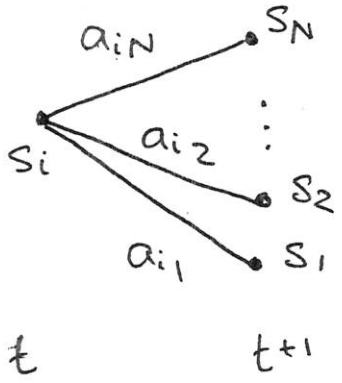
$$\beta_t(i) = P(O_{t+1}, \dots O_T \mid q_t = S_i)$$

Initialization:

$$\beta_T(i) = 1 \qquad 1 \leq i \leq N$$

Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(O_{t+1})\, \beta_{t+1}(j)$$



2. Given an HMM and an observation sequence find the single best state sequence that accounts for the observation.

Compute the probability for every possible state sequence and pick the best ?

NO!

Let

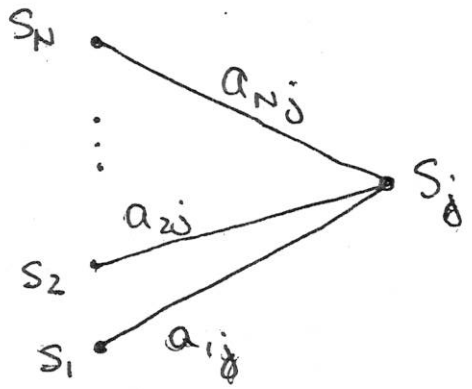
$$\delta_t(j) = \max_{q_1 \cdots q_{t-1}} P(q_1, q_2 \cdots q_t = j, O_1 \cdots O_t)$$

← over all state sequences of length $t-1$

= highest probability along the single path, which at t accounts for $O_1 \cdots O_t$ and ends at state j.

Note that

$$\delta_t(j) = \max_i \delta_{t-1}(i)\, a_{ij}\, b_j(O_t)$$



For each t and j only need to record $\delta_t(j)$ and i that maximized $\delta_{t-1}(i)\, a_{ij}$

Viterbi Algorithm

Initialization:

$$\delta_1(i) = \pi_i b_i(O_1) \qquad 1 \le i \le N$$

$$\psi_1(i) = 0$$

Recursion:

$$\delta_t(j) = \max_i \delta_{t-1}(i) a_{ij} b_j(O_t) \qquad 2 \le t \le T, \quad 1 \le j \le N$$

$$\psi_t(j) = \operatorname{argmax}_i \left[\delta_{t-1}(i) a_{ij} \right] \qquad 1 \le i \le N$$

i.e. $\psi_t(j)$ is the state (that is i) for which $\delta_{t-1}(i) a_{ij}$ is maximized

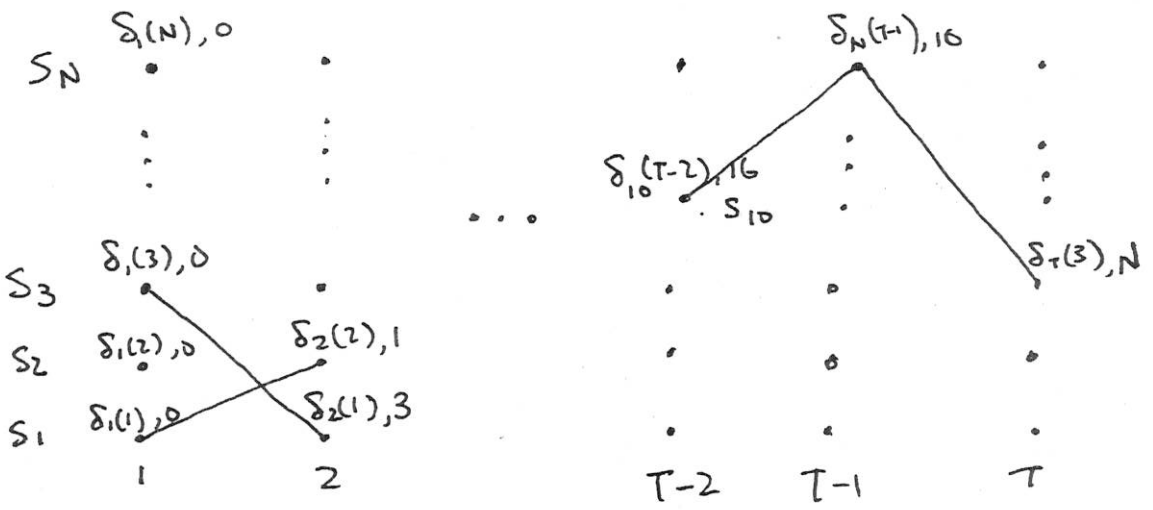
Termination:

$$p^* = \max \left[\delta_T(i) \right]$$

$$q_T^* = \operatorname{argmax}_i \left[\delta_T(i) \right]$$

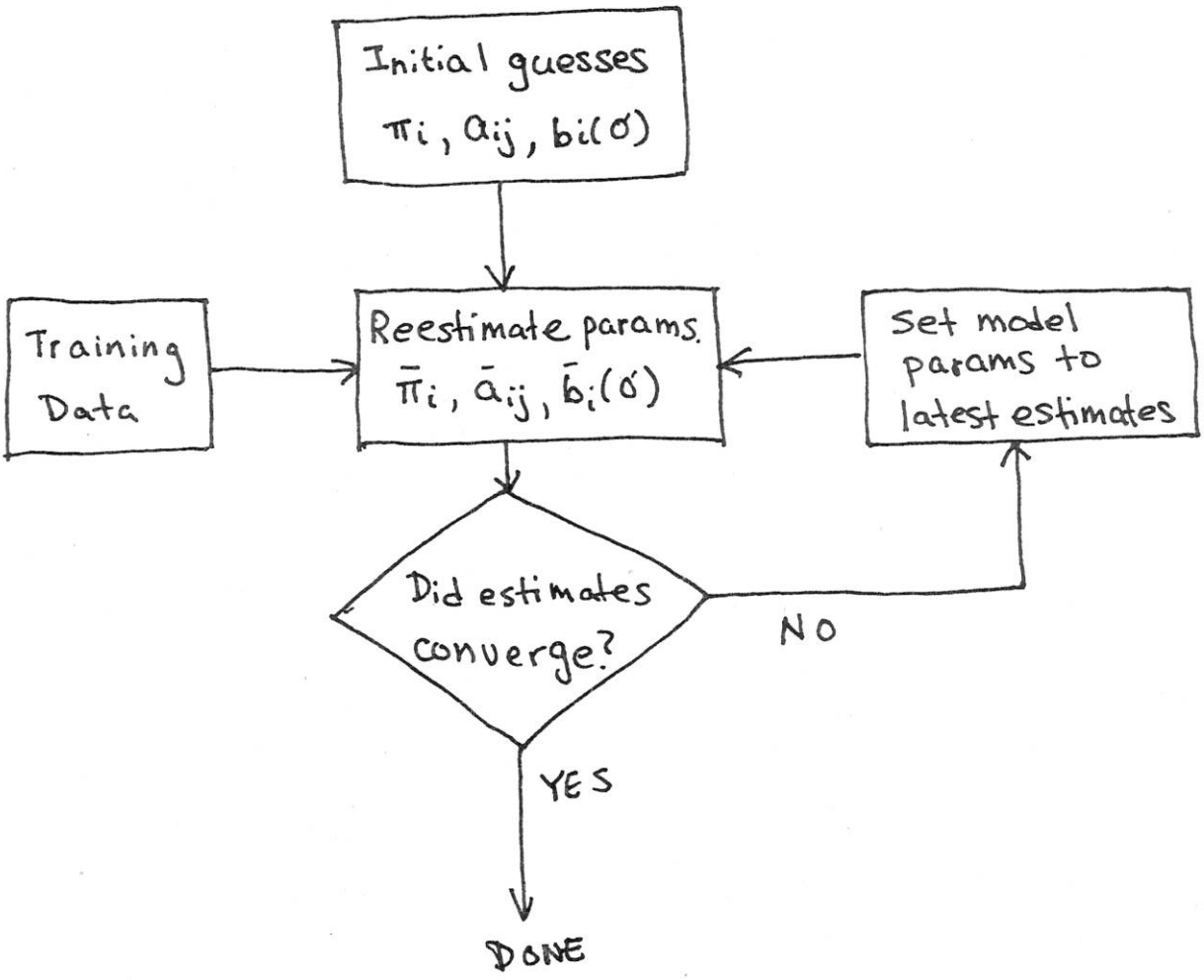
Back tracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \qquad t = T-1, T-2 \cdots, 1$$



3. Given an HMM and an observation sequence, find the parameters so that the probability of observation is maximized.

IF we could observe states, then we would only need to count state transitions, etc. But instead we need to use an iterative procedure.



Baum-Welch or forward-backward algorithm.

Reestimation formulas

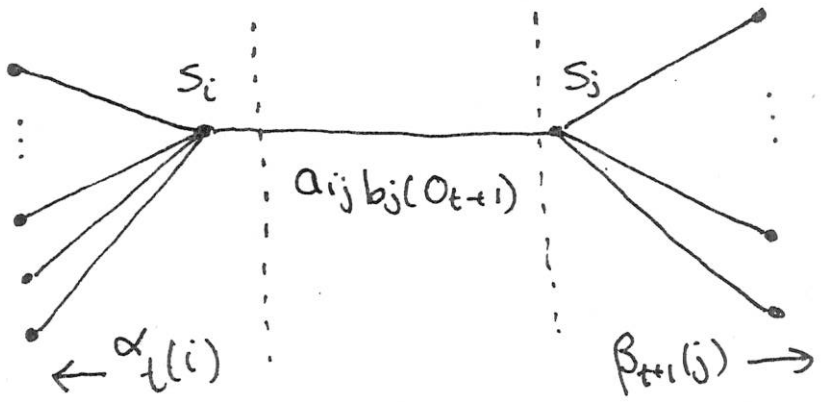
Let

- $\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid \underline{O})$ $\quad \underline{O} = O_1, O_2 \cdots O_T$

  = Prob. of being in state S_i at time t, and in state S_j at time $t+1$, given the observation sequence

  $= P(q_t = S_i, q_{t+1} = S_j, \underline{O}) / P(\underline{O})$

But $P(q_t = S_i, q_{t+1} = S_j, \underline{O}) = \alpha_t(i)\, a_{ij}\, b_j(O_{t+1})\, \beta_{t+1}(j)$



And

$$P(\underline{O}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i)\, a_{ij}\, b_j(O_{t+1})\, \beta_{t+1}(j)$$

$$= \sum_{i=1}^{N} \alpha_T(i)$$

- Let $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$ = Prob. of being in State S_i at time t, given $\underline{O}$

B-W reestimation formulas (cont'd)

We get

$\bar{\pi}_i = \gamma_1(i)$ = expected frequency in state S_i at time=1

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(ij)}{\sum_{t=1}^{T-1} \gamma_t(i)} = \frac{\text{expected \# of transitions from } S_i \text{ to } S_j}{\text{expected \# of times in } S_i}$$

* $$\bar{b}_j(k) = \frac{\sum_{\substack{t=1 \\ O_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} = \frac{\text{expected \# of times in } S_j \text{ and observing symbol } v_k}{\text{expected \# of times in } S_j}$$

* If the observation consists of real numbers and the output probability is a parametric function (e.g. Gaussian) then the parameters of the output probability density have to be estimated iteratively.

Continuous Speech Recognition (CSR)

continuous speech - string of spoken words with optional pauses between words.

What is CSR?

Let $w = (w_1, w_2 \cdots w_n)$ word sequence

$w_i \in$ vocabulary

A = acoustic observation, i.e., speech to be recognized

CSR: Find the word sequence $\hat{w}$ that maximizes $P(w|A)$.

i.e. want $\hat{w}$ such that

$$P(\hat{w}|A) = \max_w P(w|A)$$

$$P(w|A) = \frac{P(A|w)\,P(w)}{P(A)} \leftarrow \text{independent of } w$$

$\Rightarrow$ Find $\hat{w}$ such that

$$P(A|\hat{w})P(\hat{w}) = \max_w P(A|w)\,P(w)$$

Find w to maximize $P(A|w)P(w)$

$P(A|w)$ = prob. of acoustic observation given w
$\Rightarrow$ computed from Acoustic Models

$P(w)$ = prob. of word sequence w
$\Rightarrow$ computed from Language Model

How CSR is to be done in principle.

For each $w = (w_1, w_2, \cdots)$ that can be formed from the words in the vocabulary {

Compute $P(A|w)$: For every possible way of aligning w with A, compute the probability of alignment & sum all probabilities.

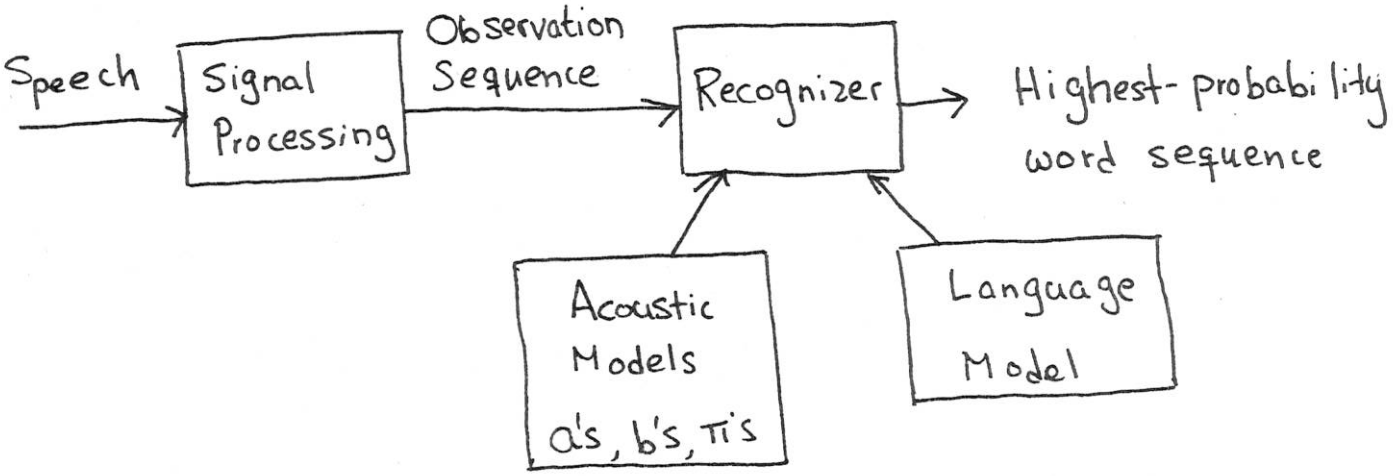
Compute $P(w)$:

}

Pick w for which $P(A|w)P(w)$ is max.

$\Rightarrow$ Need fast search algorithm

A Continuous Speech Recognizer

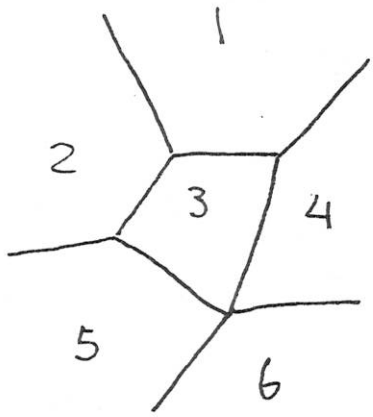


What do you use for the observation sequence?

1. Sequence of feature vectors computed at regular intervals (e.g. every 10 mseconds).

   $b_i(O_t)$ is often a continuous function of the feature vector (Also parametric.)

2. Sequence of labels (discrete symbols) obtained from the sequence of feature vectors by assigning one of M labels to each feature vector.



Example:

Divide the space of feature vectors into M cells. Label each cell. If a feature vector belongs to cell i, its label is i.

$b_i(O_t) = b_i(\text{"1"})$ if O_t = "1"

etc.

Acoustic Models for computing P(A|w)

HMM of a word sequence =

Concatenation of word HMM's

+ (Additional modeling for contextual effects at word boundaries)

popular HMM of a word = concatenation of phoneme HMM's.

(whole word HMM's are not practical for large vocabulary.)

Phoneme HMM's : Issues for model designers

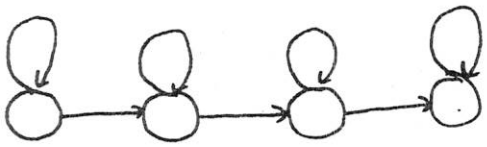
1. How many states per phoneme ?

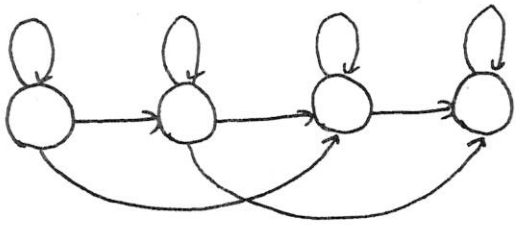
   A: A few (1, 2, 3, 4, 5 ···)

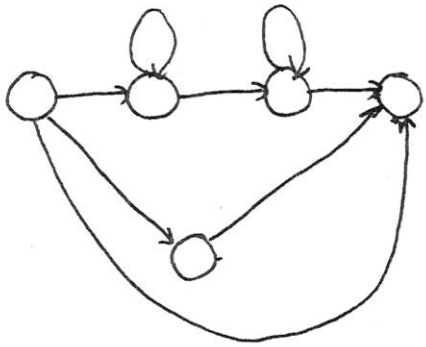
2. which model topology ?

   A: Some sort of left-to-right topology.

Examples.







3. Which feature vectors?

A: Some sort of "spectral" information
(~ 10 - 50 components)

Examples.

FFT magnitudes
LPC coefficients
Cepstral coefficients $\mathcal{F}^{-1}[\log|X(\omega)|]$ ~ time function

4. What kind of $b_j(O_t)$?

A: Discrete symbol or continuous density?

5. Context dependent phonemes? "Should there be a separate model for 'a' in 'cat' and 'a' in 'sash'?"

A: Some kind of context dependent phonemes usually used. (Possibly up to several thousand of them!) Contex includes neighboring phonemes, stress, and word boundary.

6. Speaker dependent or speaker independent models?

Language Model

n-grams are popular

$$p(w_i \mid w_{i-1}, w_{i-2}, \ldots w_{i-(n-1)})$$

$n > 3$ rarely used - need a lot of training text

Currently obtainable performance. (Examples as of Nov., 1992)

- 20,000 word vocabulary based on Wall Street Journal articles
- Speaker independent recognition
- Training & test speech consist of read sentences from the Wall Street Journal

Word error rate in the low teens.

Near real time performance possible.