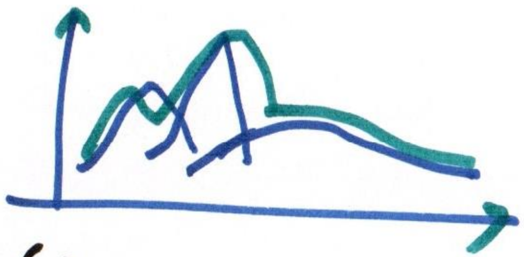
O_t ← obs vector (39-MFCC)

$P(\lambda_\ell \mid O_t)$ MAP rule

λ_ℓ ← One of L classes

$P(O_t \mid \lambda_\ell)$



$\{c_k, \mu_k, \Sigma_k\}$

c_k ← k^{th} Gaussian in GMM

Next, consider the classification of a time sequence of spectral (frame) vectors.

$$\bar{O} = (O_1, O_2, \ldots O_T) \leftarrow \text{T-frame in utterance}$$

Problem: determine the uttered word $\{\omega_\ell\}$

$$\boxed{\lambda_\ell = \arg\max P(\lambda_\ell \mid \bar{O})}$$

We use HMM (a sequence classifier)

Find $\arg\max_{\lambda_\ell} \xi P(\lambda_\ell \mid \bar{o})$

$$= \arg\max_{\lambda_\ell} P(\bar{o} \mid \lambda_\ell) \cdot P(\lambda_\ell)$$

$$\boxed{= \arg\max_{\lambda_\ell} \xi P(\bar{o} \mid \lambda_\ell)}$$ if the words are equally likely

The HMM for a word is a statistical model : $\lambda = [\bar{\pi}, A, B]$

CDEEP
IIT Bombay

EE 679 L____ / Slide 3

An HMM is a statistical model of the joint probability of a collection of random variables

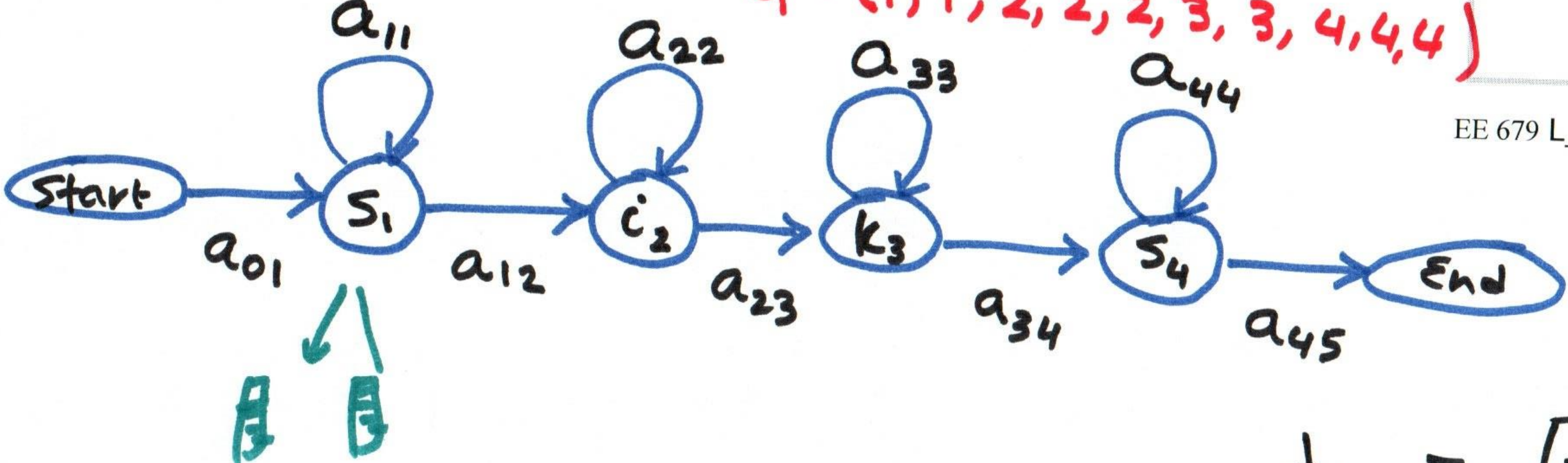
$$\{o_1, \ldots o_T, q_1, \ldots q_T\}$$

The o_t are the continuous-valued observations & the q_t are the corresp. "states" ← hidden
(discrete)

Six = s - i - k - s

Ex : $\bar{o} = o_1, o_2, \dots, o_T \rightarrow q_1, q_2 \dots q_{T=10}$

$Q = (1, 1, 2, 2, 2, 3, 3, 4, 4, 4)$



$\lambda_\ell \equiv [\pi, A, B]$

$A \downarrow a_{ij}$, $B \downarrow b_j$

π_i ("initial probabilities) $= P(q_1 = i)$

"Transition probabilities"

$$a_{ij} = P(q_t = j \mid q_{t-1} = i) , \qquad 1 \leq i, j \leq N \quad \text{(N: \# states)}$$

"Obs. probab." $b_j(o_t) = P(o_t \mid q_t = j)$

$$\sum_i a_{ij} = 1$$

We have 2 "conditional independence" assumptions

i) First-order Markov hypothesis :

the past history has no influence on the chain's evolution if the present is specified.

ii) Conditionally i.i.d. hypothesis :

given the state sequence, the observation sequence is an i.i.d. process.

i.e. the present observ^n distrib^n depends on the present state only, not on the past states or past observ^{n}s.

We want to find

$$\lambda = \arg\max_{\lambda_\ell} P(\bar{O} \mid \lambda) \cdot P(\lambda)$$

model: $\lambda_\ell = [A_\ell, B_\ell, \pi_\ell]$

CDEEP
IIT Bombay

EE 679 L ____ / Slide 6

ML classifier : $\lambda = \arg\max_{\lambda_\ell} P(\bar{O} \mid \lambda)$

We have : $P(\bar{O} \mid \lambda) = \sum_{Q} P(\bar{O}, Q \mid \lambda)$

N states
T frames
$\Rightarrow N^T$ possible state seq

But $P(\bar{O}, Q \mid \lambda) = P(\bar{O} \mid Q, \lambda) \cdot P(Q \mid \lambda)$

$$
\begin{array}{ccccc}
\text{"1"} & \text{"1"} & \text{"2"} & & \\
\uparrow & \uparrow & \uparrow & & \\
q_1 & q_2 & q_3 & \cdots & q_T \\
\downarrow & \downarrow & \downarrow & & \downarrow \\
O_1 & O_2 & O_3 & \cdots & O_T
\end{array}
$$

$$P(Q \mid \lambda) = a_{q_{T-1} q_T} \cdot \;\ldots\; \cdot a_{q_1 q_2} \pi_{q_1}$$

$$P(\bar{O} \mid Q, \lambda) = b_{q_T}(O_T) \cdot \;\ldots\; \cdot b_{q_1}(O_1)$$

$$\Rightarrow \quad P(\bar{O}, Q \mid \lambda) =$$

$$\underbrace{b_{q_T}(O_T) \cdot a_{q_{T-1} q_T}} \cdot \;\ldots\; a_{q_1 q_2} \cdot b_{q_1}(O_1)\, \pi_{q_1}$$

$$\boxed{\lambda = \underset{\lambda_l}{\arg\max} \sum_{Q} P(\bar{O}, Q \mid \lambda)} \quad \Rightarrow \text{complexity} \sim O(T \cdot N^T)$$

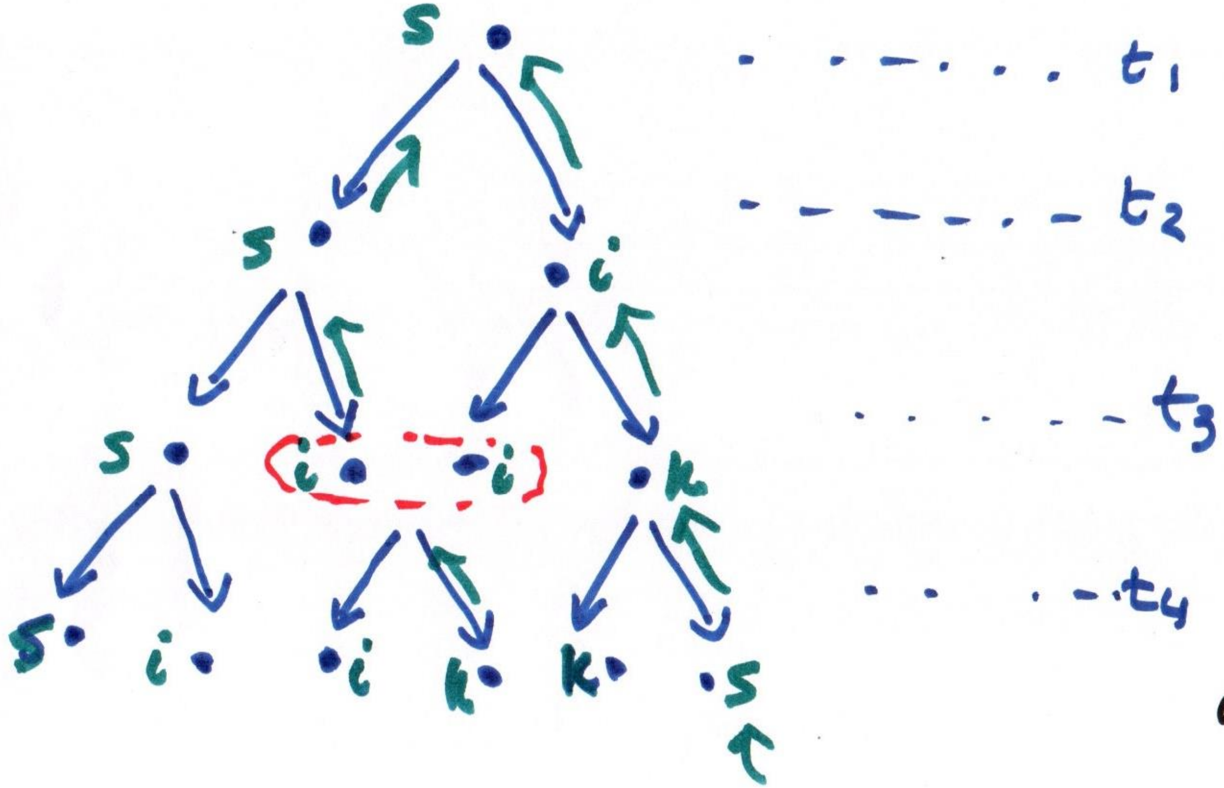
Approximate ML classific^n with the Viterbi algorithm

We do not compute $P(\bar{o}|\lambda)$ over all possible state sequences, but only for the best state sequence.

$$\lambda = \arg\max_{\lambda_i} \left(\max_{Q} P(\bar{o}, Q \mid \lambda) \right)$$

The Viterbi algo finds the most likely state seq given an HMM by finding a min. cost path through a state-time trellis graph.

Ex : s - i - k - s



Best forward prob. in state j upto time t

$\alpha_t(j) =$

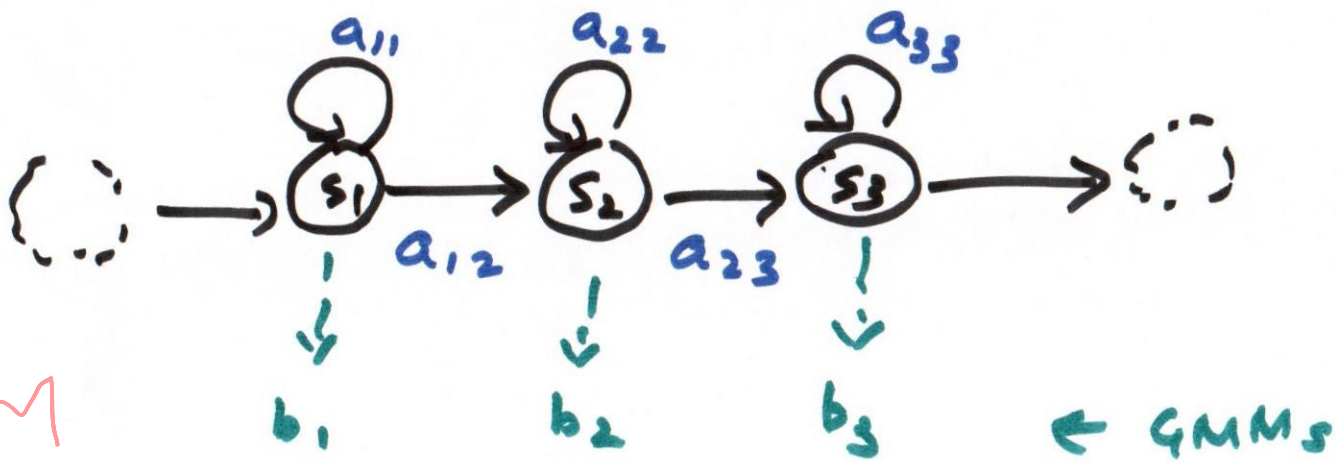
$\sum_i \alpha_{t-1}(i)\, a_{ij}\, b_j(o_t)$

$\sim O(N^2 T)$

More practical for large vocabulary is

Phone-based recognition

A basic phone model → 3-state HMM



A word HMM can be got by cascading phone HMMs

$\lambda \longrightarrow \omega \longrightarrow W$

phone word sentence (seq of words)