Introduction to Hidden Markov Models

Peter Bell

Automatic Speech Recognition— ASR Lecture 4
23 January 2025

Overview

HMMs

- Introduction to hidden Markov models
- HMMs for ASR
- Likelihood computation with the forward algorithm

Fundamental Equation of Statistical Speech Recognition

If X is the sequence of acoustic feature vectors (observations) and W denotes a word sequence, the most likely word sequence W^* is given by

$$W^* = \arg\max_{W} P(W \mid X)$$

Fundamental Equation of Statistical Speech Recognition

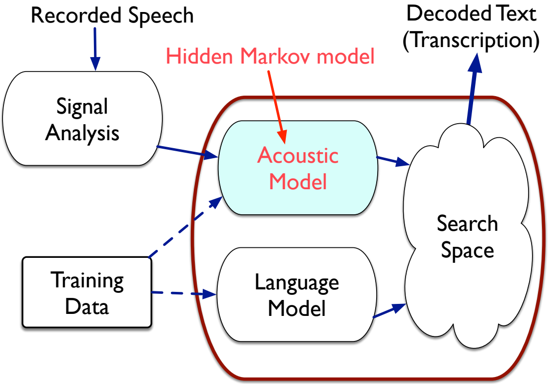
If X is the sequence of acoustic feature vectors (observations) and W denotes a word sequence, the most likely word sequence W* is given by

$$\mathrm{W}^* = \arg\max_{\mathrm{W}} P(\mathrm{W} \mid \mathrm{X})$$

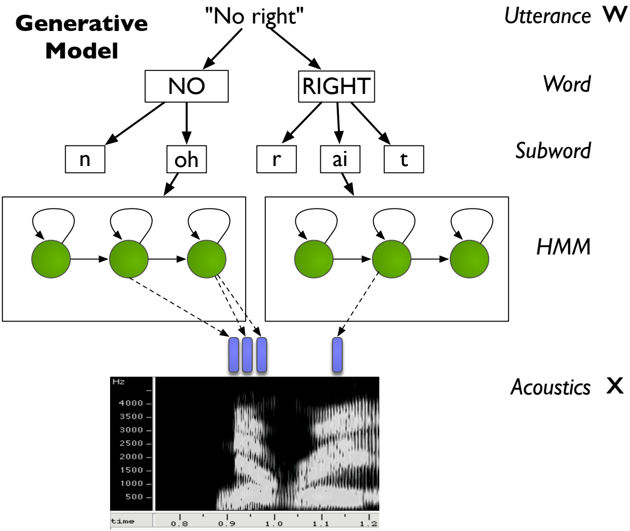
Applying Bayes' Theorem:

$$\begin{aligned} P(\mathrm{W} \mid \mathrm{X}) &= \frac{p(\mathrm{X} \mid \mathrm{W}) P(\mathrm{W})}{p(\mathrm{X})} \\ &\propto p(\mathrm{X} \mid \mathrm{W}) P(\mathrm{W}) \\ \mathrm{W}^* &= \arg\max_{\mathrm{W}} \underbrace{p(\mathrm{X} \mid \mathrm{W})}_{\substack{\text{Acoustic} \\ \text{model}}} \quad \underbrace{P(\mathrm{W})}_{\substack{\text{Language} \\ \text{model}}} \end{aligned}$$

Acoustic Modelling



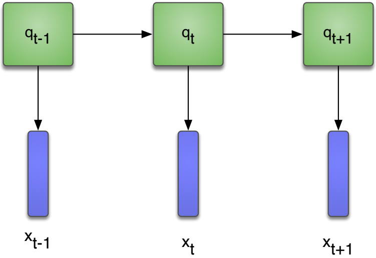
Hierarchical modelling of speech



The Hidden Markov model

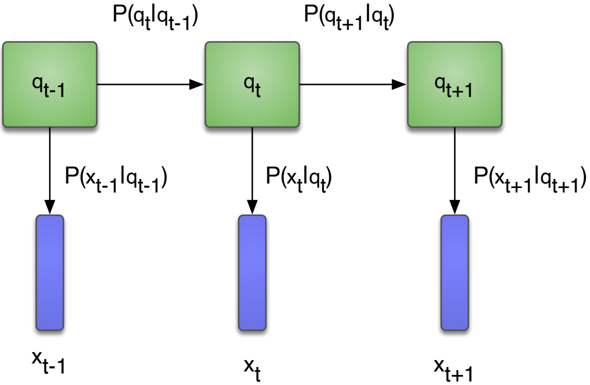
- A statistical model for time series data with a set of **discrete** states $\{1, \ldots, J\}$ (we index them by j or k)
- At each time step t:
  - the model is in a fixed state q_t.
  - the model generates an observation, x_t, according to a probability distribution that is specific to the state
- We don't actually observe which state the model is in at each time step – hence **"hidden"**.
- Observations can be either continous or discrete (usually the former)

HMM probabilities



- Imagine we know the state at a given time step t, $q_t = k$
- Then the probability of being in a new state, j at the next time step, is dependent only on q_t. This is the **Markov** assumption.
- Alternatively: q_{t+1} is *conditionally independent* of $q_1, \ldots, q_{t-1}$, given q_t.

HMM assumptions



Observation x_t is *conditionally independent* of other observations, given the state that generated it, q_t

HMM parameters

The parameters of the model, λ, are given by:

- Transition probabilities $a_{kj} = P(q_{t+1} = j | q_t = k)$
- Observation probabilities $b_j(\mathrm{x}) = P(\mathrm{x} | q = j)$

HMM topologies

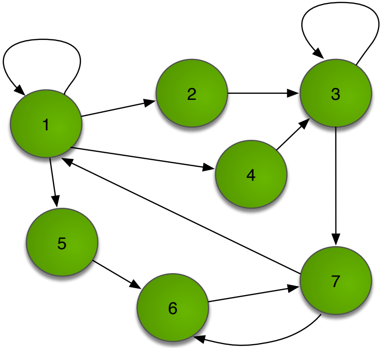
- The HMM topology determines the set of allowed transitions between states

HMM topologies

- The HMM topology determines the set of allowed transitions between states
- In principle any topology is possible

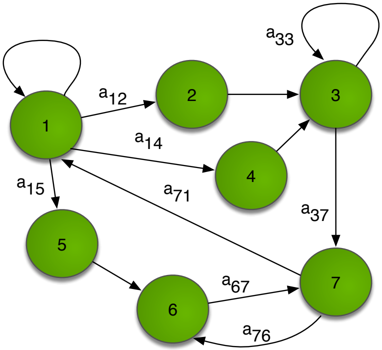
HMM topologies

- The HMM topology determines the set of allowed transitions between states
- In principle any topology is possible



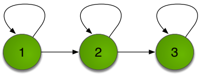
HMM topologies

- The HMM topology determines the set of allowed transitions between states
- In principle any topology is possible



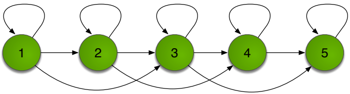
Not all transition probabilities are shown

Example topologies



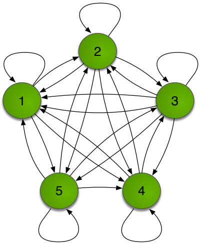
left-to-right model

$$\begin{pmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix}$$



parallel path left-to-right model

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 \\ 0 & a_{22} & a_{23} & a_{24} & 0 \\ 0 & 0 & a_{33} & a_{34} & a_{35} \\ 0 & 0 & 0 & a_{44} & a_{45} \\ 0 & 0 & 0 & 0 & a_{55} \end{pmatrix}$$



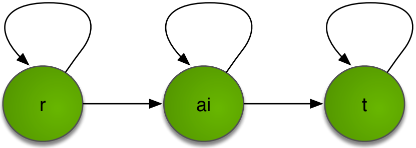
ergodic model

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{pmatrix}$$

Traditional speech recognition: left-to-right HMM with $3 \sim 5$ states
Speaker recognition: ergodic HMM

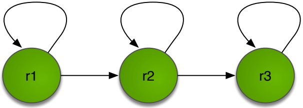
HMMs for ASR

We generally model words or phones with a left-to-right topology with self loops.



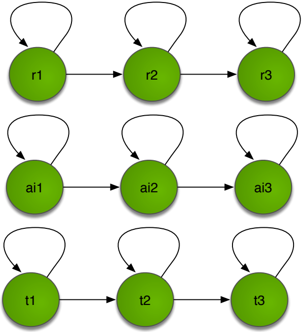
HMMs for ASR

Traditional HMMs for ASR tend to model each phone with three distinct states (this also enforces a minimum phone duration of three frames of observations)



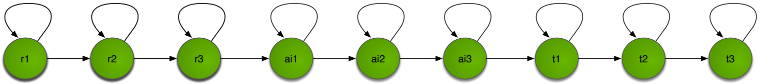
HMMs for ASR

Traditional HMMs for ASR tend to model each phone with three distinct states (this also enforces a minimum phone duration of three frames of observations)



HMMs for ASR

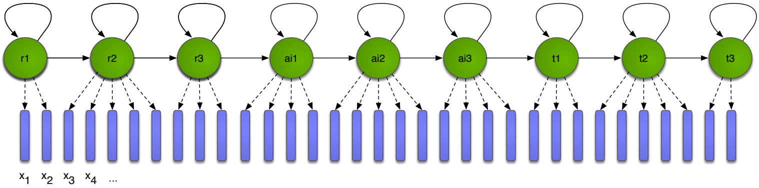
Traditional HMMs for ASR tend to model each phone with three distinct states (this also enforces a minimum phone duration of three frames of observations)



The phone model topologies can be concatenated to form a HMM for the whole word

HMMs for ASR

Traditional HMMs for ASR tend to model each phone with three distinct states (this also enforces a minimum phone duration of three frames of observations)



This model naturally generates an alignment between states and observations (and hence words/phones).

Computing likelihoods with the HMM

Suppose we have a sequence of observations of length T, $X = (x_1, \ldots, x_T)$, and Q is a known state sequence, $(q_1, \ldots, q_T)$. Then we can use the HMM to compute the joint likelihood of X and Q:

$$P(X, Q; \lambda) = P(q_1)P(x_1|q_1)P(q_2|q_1)P(x_2|q_2)\ldots \quad (1)$$

$$= P(q_1)P(x_1|q_1)\prod_{t=2}^{T} P(q_t|q_{t-1})P(x_t|q_t) \quad (2)$$

$P(q_1)$ denotes the initial occupancy probability of each state

The three problems of HMMs

Working with HMMs requires the solution of three problems:

1. **Likelihood** Determine the overall likelihood of an observation sequence $X = (x_1, \ldots, x_t, \ldots, x_T)$ being generated by a known HMM topology, $\mathcal{M}$.

The three problems of HMMs

Working with HMMs requires the solution of three problems:

1. **Likelihood** Determine the overall likelihood of an observation sequence $X = (x_1, \ldots, x_t, \ldots, x_T)$ being generated by a known HMM topology, $\mathcal{M}$.
2. **Decoding and alignment** Given an observation sequence and an HMM, determine the most probable hidden state sequence

The three problems of HMMs

Working with HMMs requires the solution of three problems:

1. **Likelihood** Determine the overall likelihood of an observation sequence $X = (x_1, \ldots, x_t, \ldots, x_T)$ being generated by a known HMM topology, $\mathcal{M}$.
2. **Decoding and alignment** Given an observation sequence and an HMM, determine the most probable hidden state sequence
3. **Training** Given an observation sequence and an HMM, find the state occupation probabilities

Computing likelihood

1. **Likelihood** Determine the overall likelihood of an observation sequence $X = (x_1, \ldots, x_t, \ldots, x_T)$ being generated by a known HMM topology, $\mathcal{M}$.
   $\rightarrow$ the *forward algorithm*

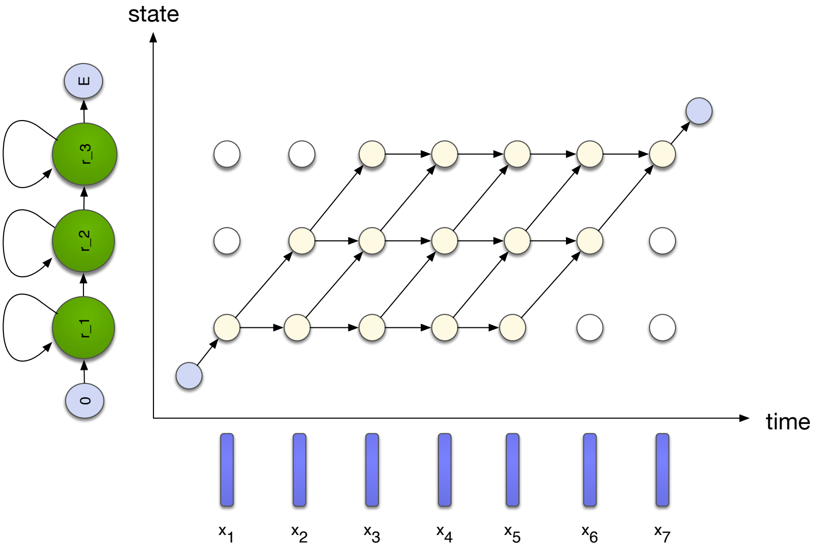
NB. We do **not** know the state sequence!

Notes on the HMM topology

By talking about HMM topologies in the context of speech recognition, $\mathcal{M}$, we can mean:

- A restricted left-to-right topology based on a known word/sentence, leading to a "trellis-like" structure over time
- A much less restricted topology based on a grammar or language model – or something in between
- Some algorithms are not (generally) suitable for unrestricted topologies

Example: trellis for a 3-state left-to-right phone HMM



Likelihood

- Goal: determine $p(\mathsf{X} \mid \mathcal{M})$

Likelihood

- Goal: determine $p(\mathrm{X} \mid \mathcal{M})$
- Sum over all possible state sequences $Q = (q_1, \ldots, q_T)$ that could result in the observation sequence $\boldsymbol{X}$

$$p(\mathrm{X}|\mathcal{M}) = \sum_{Q \in \mathcal{Q}} P(\mathrm{X}, Q|\mathcal{M})$$

$$= \sum_{Q \in \mathcal{Q}} P(q_1) P(\mathrm{x}_1|q_1) \prod_{t=2}^{T} P(q_t|q_{t-1}) P(\mathrm{x}_t|q_t)$$

Likelihood

- Goal: determine $p(\mathrm{X}|\mathcal{M})$
- Sum over all possible state sequences $Q = (q_1, \ldots, q_T)$ that could result in the observation sequence $\boldsymbol{X}$

$$p(\mathrm{X}|\mathcal{M}) = \sum_{Q \in \mathcal{Q}} P(\mathrm{X}, Q|\mathcal{M})$$

$$= \sum_{Q \in \mathcal{Q}} P(q_1)P(\mathrm{x}_1|q_1) \prod_{t=2}^{T} P(q_t|q_{t-1})P(\mathrm{x}_t|q_t)$$

- How many paths Q do we have to calculate?

$$\sim \underbrace{N \times N \times \cdots N}_{T \text{ times}} = N^T$$

N : number of HMM states
T : length of observation

e.g. $N^T \approx 10^{10}$ for $N=3,\ T=20$

Likelihood

- Goal: determine $p(\mathsf{X}|\mathcal{M})$
- Sum over all possible state sequences $Q = (q_1, \ldots, q_T)$ that could result in the observation sequence $\boldsymbol{X}$

$$p(\mathsf{X}|\mathcal{M}) = \sum_{Q \in \mathcal{Q}} P(\mathsf{X}, Q|\mathcal{M})$$

$$= \sum_{Q \in \mathcal{Q}} P(q_1)P(\mathsf{x}_1|q_1) \prod_{t=2}^{T} P(q_t|q_{t-1})P(\mathsf{x}_t|q_t)$$

- How many paths Q do we have to calculate?

$$\sim \underbrace{N \times N \times \cdots N}_{T \text{ times}} = N^T$$

N : number of HMM states

T : length of observation

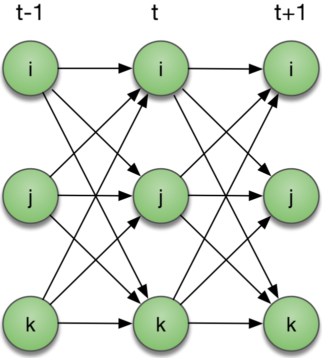
e.g. $N^T \approx 10^{10}$ for $N = 3$, $T = 20$

- Computation complexity of multiplication: $O(2TN^T)$

Likelihood: The Forward algorithm

The **Forward algorithm**:

- Rather than enumerating each sequence, compute the probabilities recursively (exploiting the Markov assumption)
- Reduces the computational complexity to $O(TN^2)$
- State time trellis for an arbitrary HMM topology



The forward probability

Define the *Forward probability*, $\alpha_j(t)$: the probability of observing the observation sequence $x_1 \ldots x_t$ and being in state j at time t:

$$\alpha_j(t) = p(x_1, \ldots, x_t, q_t = j \mid \mathcal{M})$$

We can *recursively* compute this probability

Initial and final state probabilities

It what follows it is convenient to define:

- an additional single initial state $S_I = 0$, with transition probabilities

$$a_{0j} = P(q_1 = j)$$

  denoting the probability of starting in state j
- a single final state, S_E, with transition probabilities a_{jE} denoting the probability of the model terminating in state j.
- S_I and S_E are both *non-emitting*

Likelihood: The Forward recursion

- Initialisation

$$\alpha_j(0) = 1 \qquad j = 0$$
$$\alpha_j(0) = 0 \qquad j \neq 0$$

- Recursion

$$\alpha_j(t) = \sum_{i=0}^{J} \alpha_i(t-1) a_{ij} b_j(\mathrm{x}_t) \qquad 1 \leq j \leq J,\ 1 \leq t \leq T$$

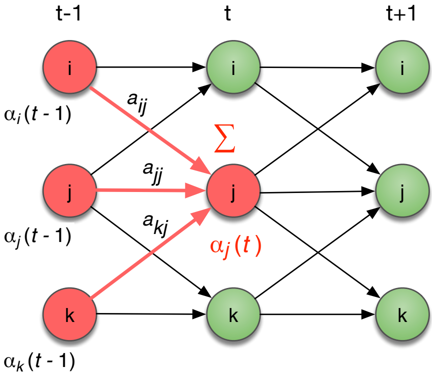
- Termination

$$p(\mathrm{X} \mid \mathcal{M}) = \alpha_E = \sum_{i=1}^{J} \alpha_i(T) a_{iE}$$

s_I: initial state, s_E: final state

Likelihood: Forward Recursion

$$\alpha_j(t) = p(x_1, \ldots, x_t, q_t = j|\mathcal{M}) = \sum_{i=1}^{J} \alpha_i(t-1) a_{ij} b_j(x_t)$$



Next

More HMM algorithms

- Finding the most likely path with the Viterbi algorithm
- Parameter estimation:
    - the Forward-Backward algorithm
    - the Expectation-Maximisation algorithm

References: HMMs

* Rabiner and Juang (1986). "An introduction to hidden Markov models", *IEEE ASSP Magazine*, **3** (1), 4–16.

  Jurafsky and Martin (2008). *Speech and Language Processing* (2nd ed.): sections 6.1–6.5; 9.2; 9.4.

  Renals and Hain (2010). "Speech Recognition", *Computational Linguistics and Natural Language Processing Handbook*, Clark, Fox and Lappin (eds.), Blackwells: sections 2.1 and 2.2.