Graphical Models for Automatic Speech Recognition

Advanced Signal Processing SE 2, SS05

Stefan Petrik

Signal Processing and Speech Communication Laboratory

Graz University of Technology

Overview

- Introduction to ASR
- Pronunciation Modeling
- Language Modeling
- Basic Speech Models
- Advanced Speech Models
- Summary

Introduction to ASR

- Find most likely sequence of words w^*, given observations X

$$w^* = \arg\max_w(P(w|X)) = \arg\max_w \frac{P(w) \cdot P(X|w)}{P(X)}$$

  - $w = w_1 ... w_m$: sequence of words
  - X : feature vectors
  - $P(w)$: language model
  - $P(X|w)$: acoustic model

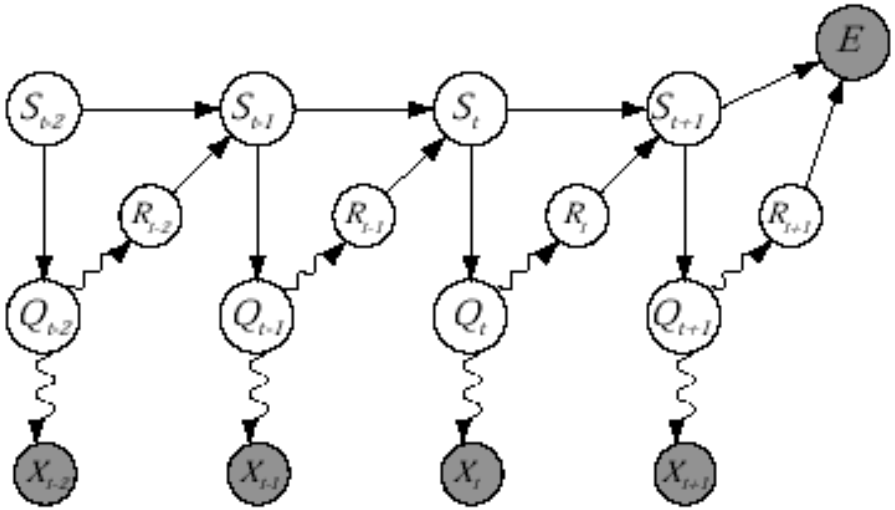
- Tasks in ASR:
  - acoustic modeling
  - pronunciation modeling
  - language modeling

Pronunciation Modeling

- Map base-forms (word dictionary based pronunciations) to surface forms (actual pronunciations)
- Use 1^{st} order Markov chain for representation
- Phones are shared across multiple words: /b/ae/g/ $\leftrightarrow$ /b/ae/t/
- Solution 1: Expanded model
  - increase state space of Q_t, to model not only phone but also position in word
  - condition on word W_t and sequential position of phone S_t : $P(Q_t|W_t, S_t)$

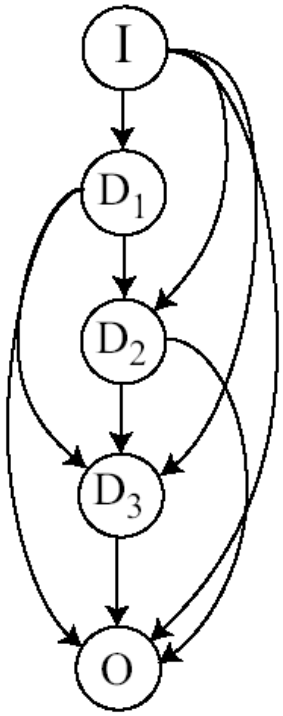
Pronunciation Modeling

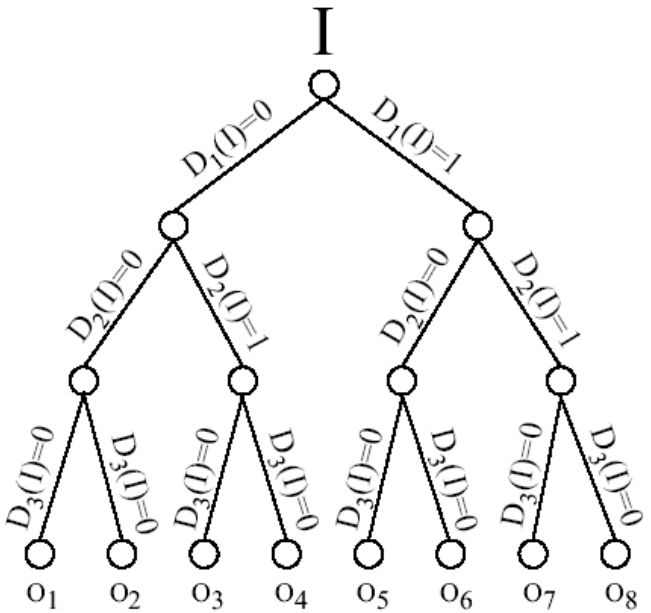
- Solution 2: Parameter-tied model
  - avoids expanded state space by parameter tying and sequence control
  - $p(S_{t+1} = i | R_t, S_t) = \delta_{i, f(R_t, S_t)}$ $\quad S_{t+1} = \begin{cases} S_t + 1 & \text{if} \quad R_t = 1 \\ S_t & \text{else} \end{cases}$



Pronunciation Modeling

- Solution 3: Decision trees
  - input node I, output node O, decision RVs R_i
  - $P(D_l = i|I) = \delta_{i,f_l(I,d_{1:l-1})}$ with decisions $d_l = f_l(I, d_{1:l-1})$





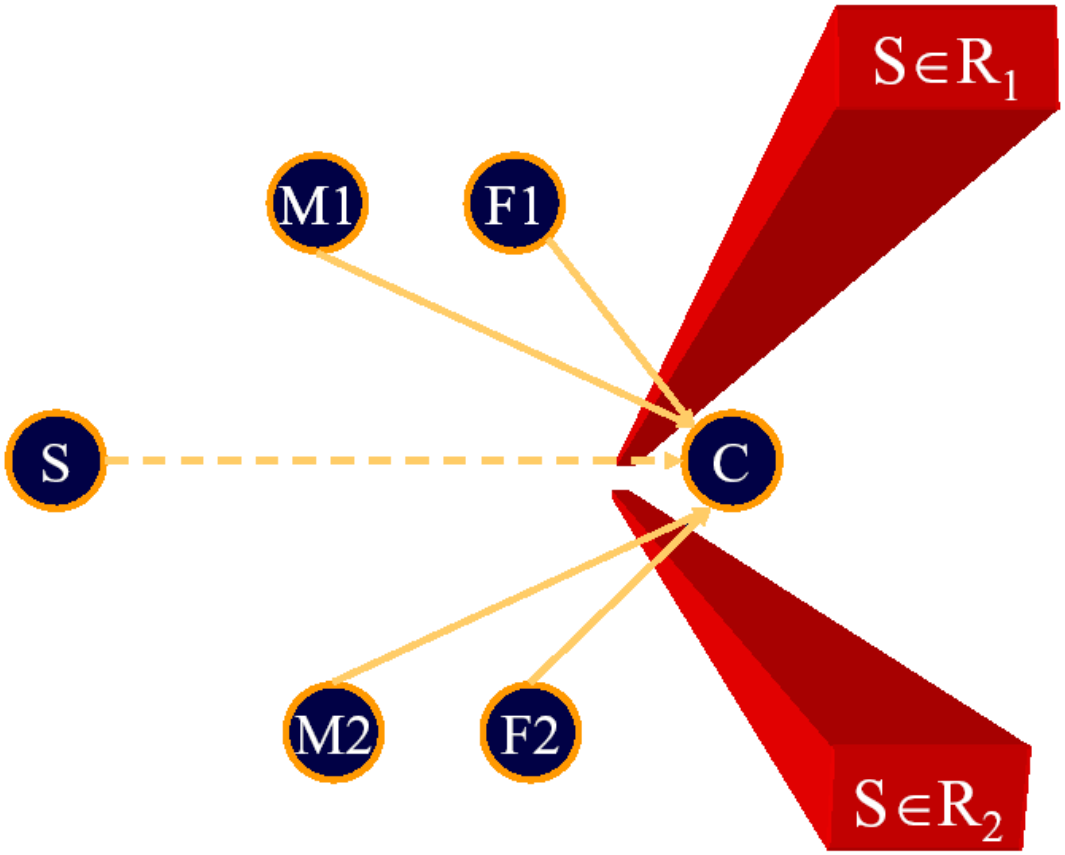
Language Modeling

- Predict next word from history of previous words
- Joint distribution: $p(W_{1:T}) = \prod_{t=1}^{n} p(W_t|W_{1:t-1})$
- Restrict to history of last $n-1$ words: $p(W_t|W_{t-n+1:t-1}) = p(W_t|H_t)$
- Problem: sparse data
- Solution: smoothing

$$\begin{aligned} p(w_t|w_{t-1},w_{t-2}) &= \alpha_3(w_{t-1},w_{t-2})f(w_t|w_{t-1},w_{t-2}) \\ &+ \alpha_2(w_{t-1},w_{t-2})f(w_t|w_{t-1}) \\ &+ \alpha_1(w_{t-1},w_{t-2})f(w_t) \end{aligned}$$

n-Grams

- Switching parents: value-specific conditional independence

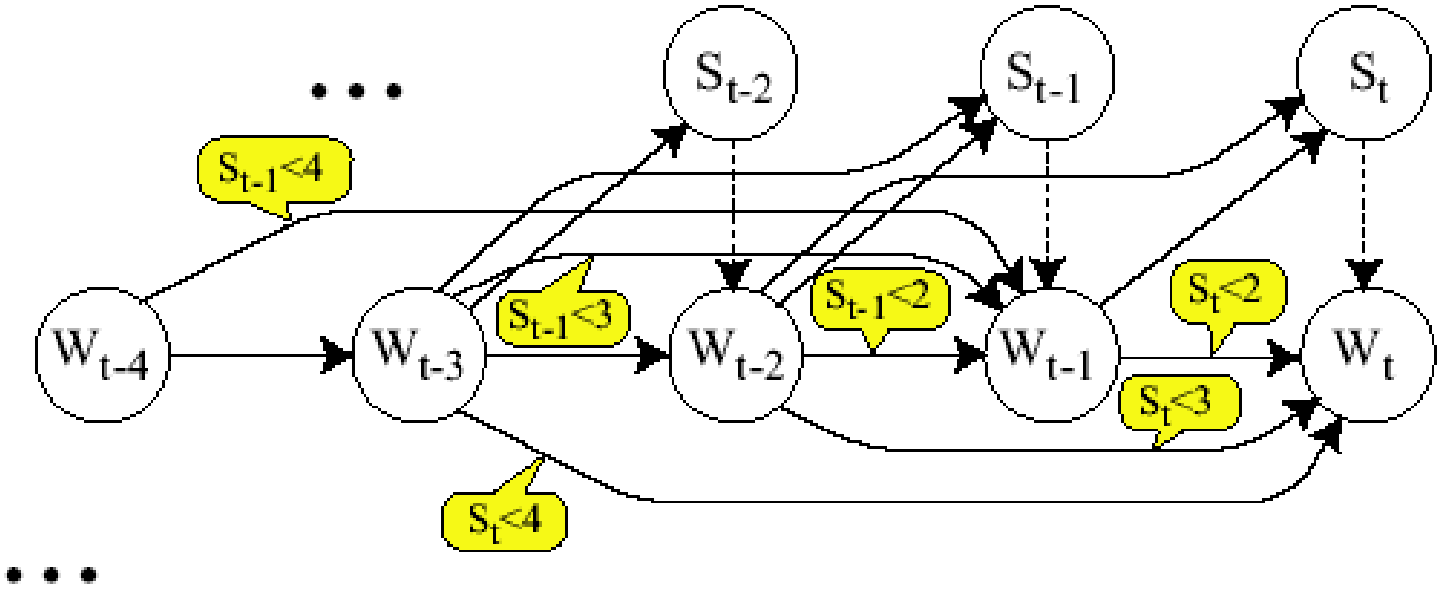


$$P(C|M1, F1, M2, F2) = \sum_i P(C|M_i, F_i, S = i)P(S \in R_i)$$

n-Grams

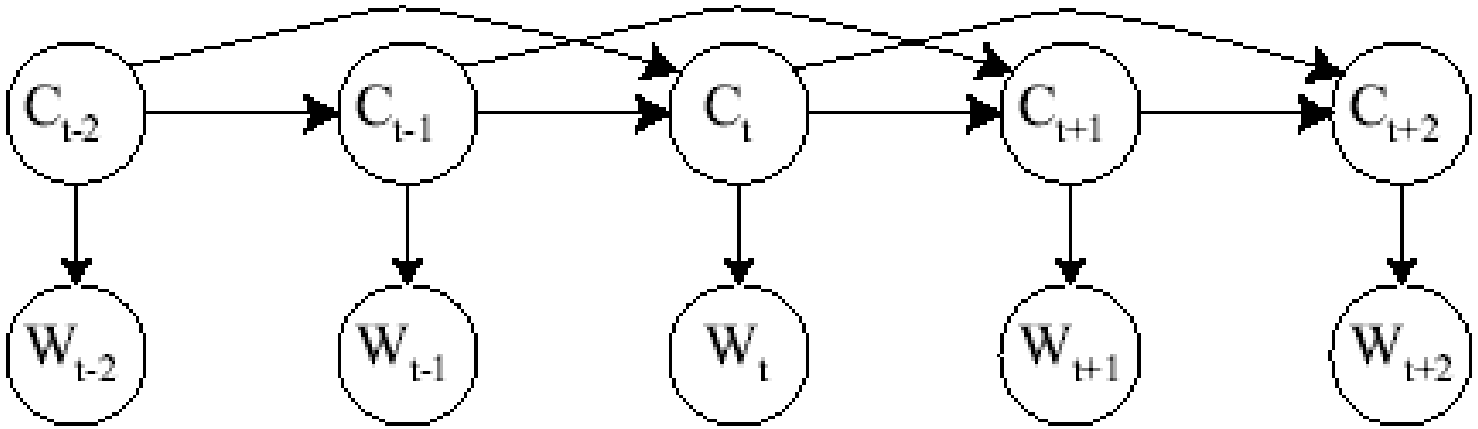
- Resulting model:

$$
\begin{aligned}
p(w_t|w_{t-1},w_{t-2}) &= \alpha_3(w_{t-1},w_{t-2})f(w_t|w_{t-1},w_{t-2}) \\
&+ \alpha_2(w_{t-1},w_{t-2})f(w_t|w_{t-1}) \\
&+ \alpha_1(w_{t-1},w_{t-2})f(w_t)
\end{aligned}
$$



Class-Based Language Model

- Idea: cluster words together and form a Markov chain over the groups
- Much lower dimensional class variables C_i than high-dimensional word variables W_i
- Syntactic, semantic or pragmatic grouping:
  - parts-of-speech: nouns, verbs, adjectives, determiners, ...
  - numerals, colors, sizes, physical values, ...
  - animals, plants, vegetables, people, ...



Class-Based Language Model

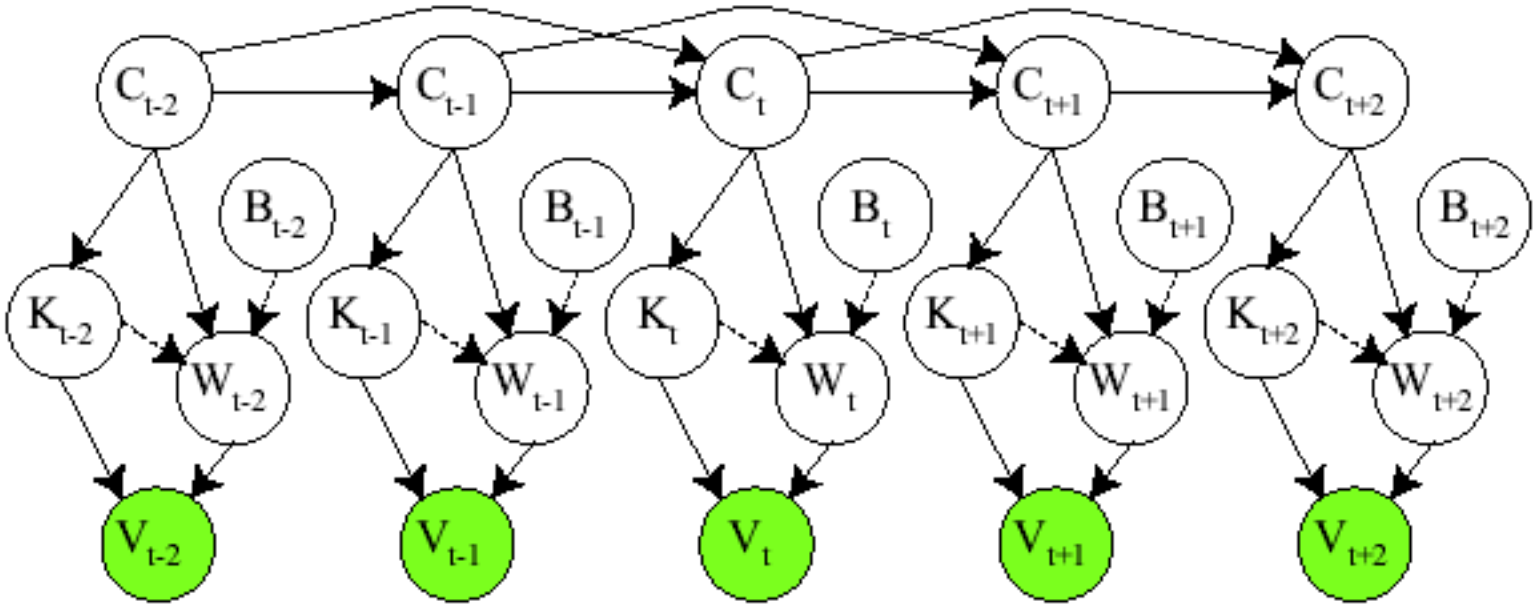
- Introduce token unk with non-zero probability for unknown words
- Vocabulary $\mathcal{W} = \{unk\} \cup \mathcal{S} \cup \mathcal{M}$ with $p_{ml}(w \in \mathcal{W}) = \frac{N(w)}{N}$
- Constraint: $p(\text{unk}) = 0.5 * p_{ml}(\mathcal{S}) = 0.5 * \sum_{w \in \mathcal{S}} p_{ml}(w)$
- Resulting probability model:

$$p_d(w) = \begin{cases} 0.5 p_{ml}(\mathcal{S}) & \text{if} \quad w = unk \\ 0.5 p_{ml}(w) & \text{if} \quad w \in \mathcal{S} \\ p_{ml}(w) & \text{otherwise} \end{cases}$$

- Condition on current class: $p_d(w|c)$

Class-Based Language Model

- Graphical model:
  - Additional observed variable V_t which is always $V_t = 1$
  - K_t, B_t : switching parents
  - C_t : word class W_t : word
- Show $p(w_t, V_t = 1|c_t) = p_d(w_t|c_t)$



Class-Based Language Model

- Conditional distributions:

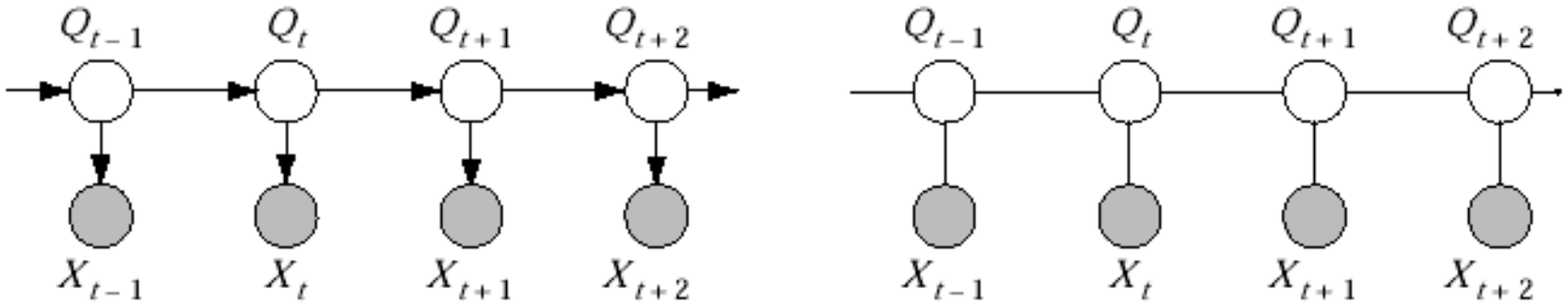
$$p(k_t|c_t) = \begin{cases} p(\mathcal{S}|c_t) & \text{if} \quad k_t = 1 \\ 1 - p(\mathcal{S}|c_t) & \text{otherwise} \end{cases} \qquad p(B_t = 0) = p(B_t = 1) = 0.5$$

$$p_M(w|c) = \begin{cases} \frac{p_m l(w|c)}{p(\mathcal{M}|c)} & \text{if} \quad w \in \mathcal{M} \\ 0 & \text{otherwise} \end{cases} \qquad p_S(w|c) = \begin{cases} \frac{p_m l(w|c)}{p(\mathcal{S}|c)} & \text{if} \quad w \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$$

$$p(w_t|k_t, b_t, c_t) = \begin{cases} p_M(w_t|c_t) & \text{if} \quad k_t = 0 \\ p_S(w_t|c_t) & \text{if} \quad k_t = 1 \quad \text{and} \quad b_t = 1 \\ \delta_{\{w_t = unk\}} & \text{if} \quad k_t = 1 \quad \text{and} \quad b_t = 0 \end{cases}$$

Basic Speech Models

- Hidden Markov Model (HMM):
  - encompasses acoustic, pronunciation, and language modeling
  - hidden chain corresponds to seq. of words, phones and sub-phones
  - hidden states $Q_{1:T}$ and observations $X_{1:T}$
  - $Q_{t:T} \perp Q_{1:t-2} | Q_{t-1}$ and $X_t \perp \{Q_{\neg t}, X_{\neg t}\} | Q_t$



  - either DGM or UGM: moralizing the graph introduces no new edges and result is already triangulated

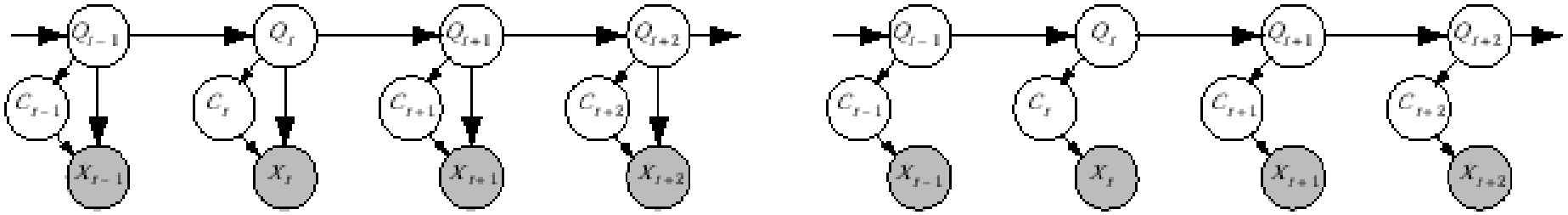
Basic Speech Models

- HMMs with mixture-of-Gaussians output:
  - explicit modeling of mixture variable

$$p(X_t|Q_t = q, C_t = i) = \mathcal{N}(x_t; \mu_{q,i}, \Sigma_{q,i}) \qquad p(C_t = i|Q_t = q) = C(q, i)$$

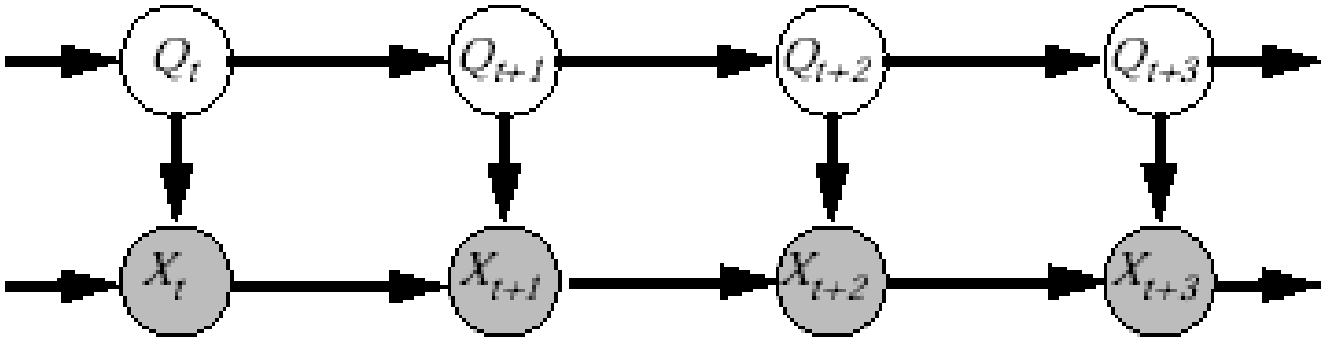
- Semi-continuous HMMs:
  - single, global pool of Gaussians, each state corresponds to a particular mixture over the pool

$$p(x|Q = q) = \sum_i p(C = i|Q = q)p(x|C = i)$$



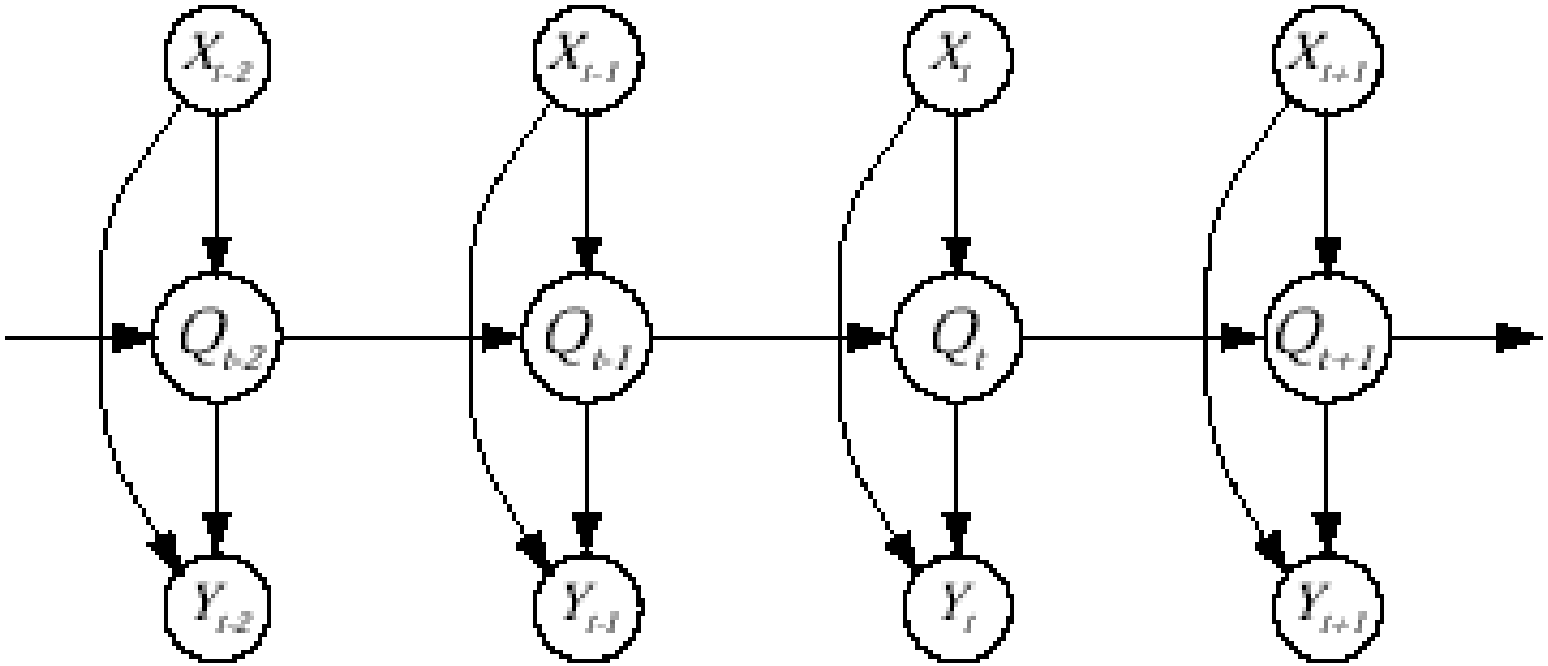
Basic Speech Models

- Auto-regressive HMMs (AR-HMM):
  - relaxes conditional independence constraint 2:
    X_{t-1} helps predicting X_t
  - result: models with higher likelihood
  - note: not to be confused with linear predicitve HMMs



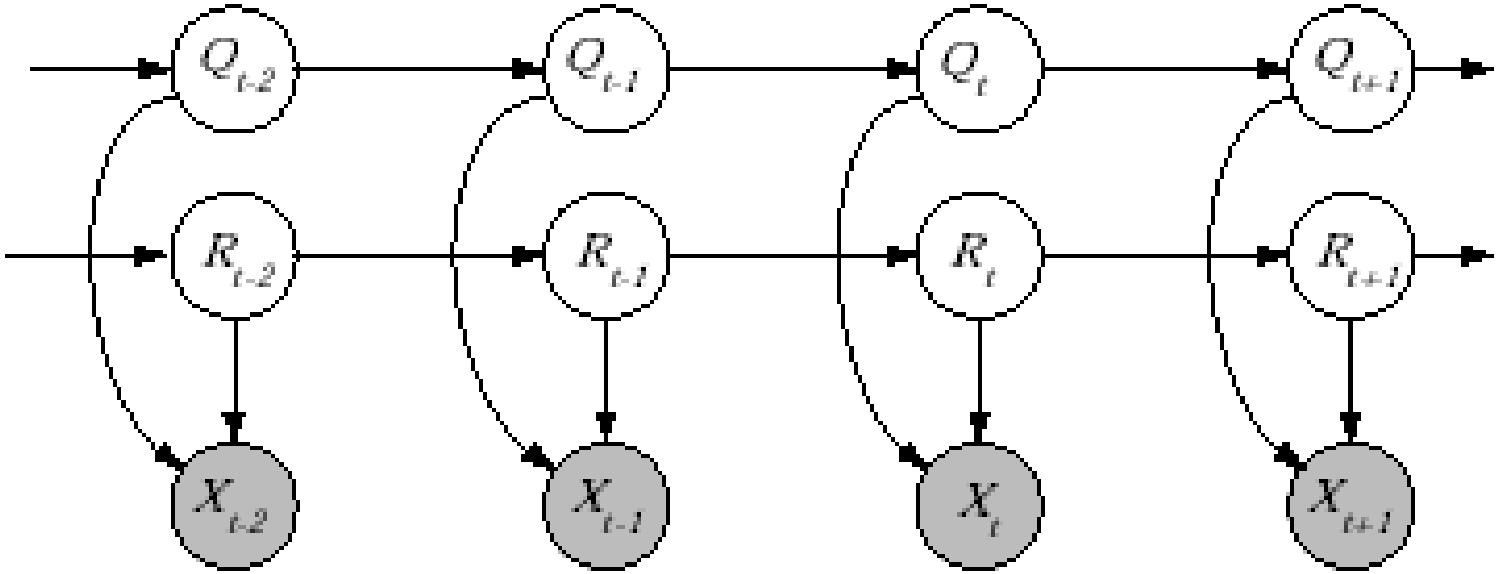
Basic Speech Models

- Input/Output HMMs (IOHMM):
  - variables corresponding to input and output at each time frame
  - given input feature stream $X_{1:T}$, try to find $E[Y|X]$
  - CPD for Q_t as 3-dim array: $P(Q_t = j|Q_{t-1} = i, X_t = k) = A(i, j, k)$
  - promising for speech enhancement



Advanced Speech Models

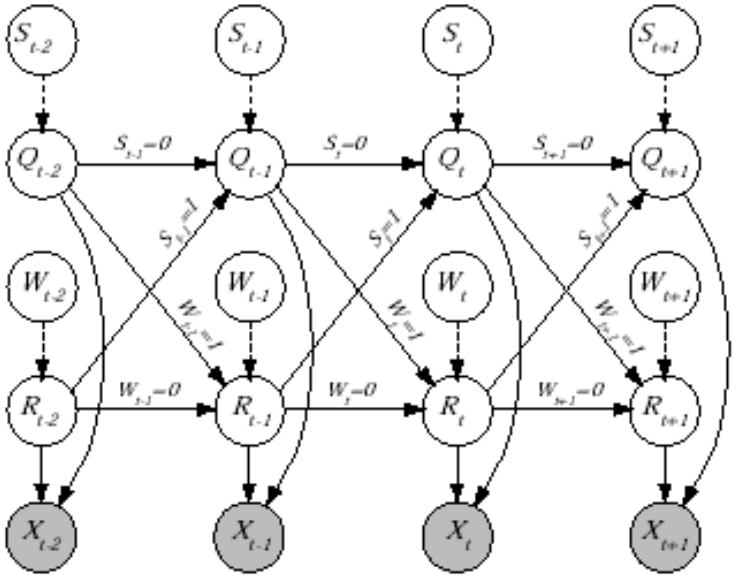
- Factorial HMMs:
  - distributed representation of the hidden state
  - special case HMM with tied parameters and state transition restrictions
  - conversion to HMM possible, but inefficient:
    complexity changes from $O(TMK^{M+1})$ to $O(TK^{2M})$



Advanced Speech Models

- Mixed-memory HMMs:
  - like factorial HMM, but fewer parameters
    (two 2-dimensional tables instead of single 3-dimensional one)
  - cond. independence: $Q_t \perp R_{t-1} | S_t = 0$ and $Q_t \perp Q_{t-1} | S_t = 1$

$$
\begin{aligned}
p(Q_t|Q_{t-1}, R_{t-1}) &= p(Q_t|Q_{t-1}, S_t = 0)P(S_t = 0) \\
&+ p(Q_t|Q_{t-1}, S_t = 1)P(S_t = 1)
\end{aligned}
$$

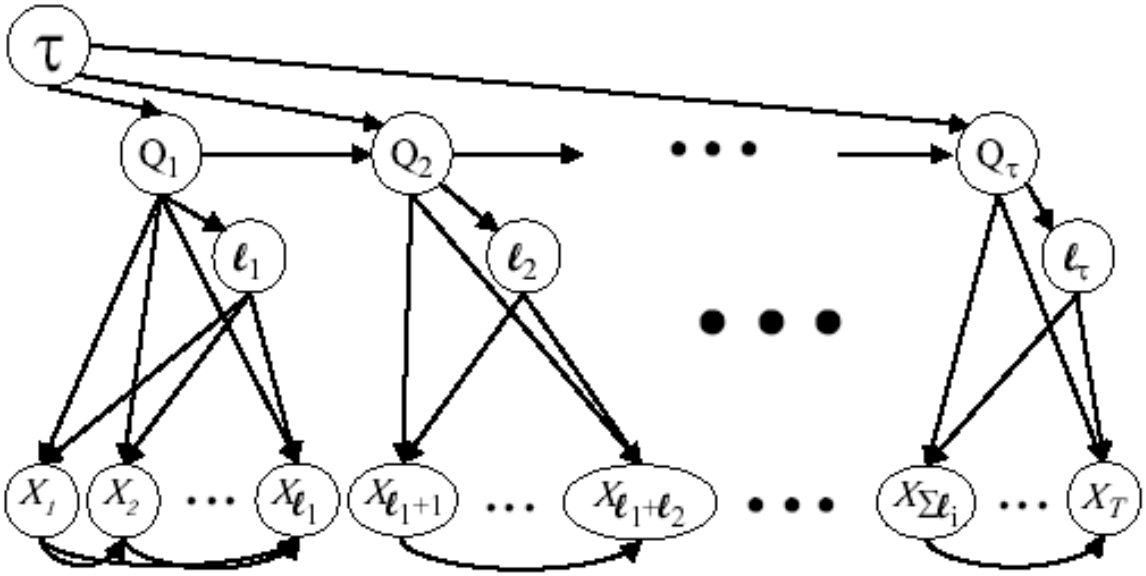


Advanced Speech Models

- Segment models:
  - each HMM state can generate sequence of observations, not just single one
  - overall joint distribution: $p(X_{1:T} = x_{1:T}) =$

    $$\sum_{\tau} \sum_{q_{1:\tau}} \sum_{l_{1:\tau}} \prod_{i=1}^{\tau} p(x_{t(i,1)}), p(x_{t(i,2)}), ..., p(x_{t(i,l_i)}), l_i | q_i, \tau) p(q_i | q_{i-1}, \tau) p(\tau)$$

  - observation segment distribution: $p(x_1, x_2, ..., x_l, l|q) = p(x_1, x_2, ..., x_l|l, q)p(l|q)$
  - plain HMM if $p(x_1, x_2, ..., x_l|l, q) = \prod_{j=1}^{l} p(x_j|q)$ and $p(l|q)$ geometric dist.



Advanced Speech Models

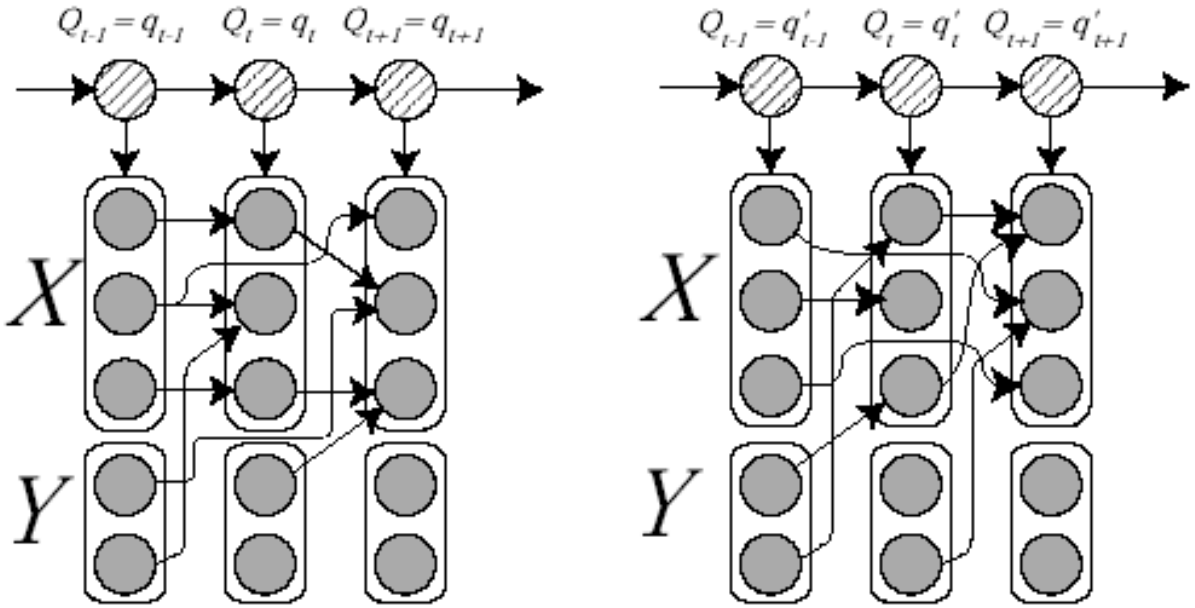
- Buried Markov Model (BMM):
  - HMM's cond. independence structure may not accurately model data
    $\Rightarrow$ additional edges between observation vectors needed
  - Idea: measure contextual information of a hidden variable
  - conditional mutual information:
    additional information $X_{<t}$ provides about X_t not already provided by Q_t

$$I(X_t; X_{<t}|Q_t) = \sum_q I(X_t; X_{<t}|Q_t = q)p(Q_t = q) = \begin{cases} > 0 & \text{add edge} \\ 0 & \text{no change} \end{cases}$$

  - underlying Markov chain in HMM is further hidden (buried) by specific cross-observation dependencies

Advanced Speech Models

- Buried Markov Model (BMM):
  - for learning, measure discriminative mutual information between X and its potential set of parents Z
  - EAR (explaining away residual): $EAR(X,Z) = I(X;Z|Q) - I(X;Z)$
  - $\arg\max_Z EAR(X,Z) \Rightarrow$ optimized posterior probability for Q



Summary

- Some well-known speech models presented in terms of graphical models
- Used for acoustic, pronunciation and language modeling
- Standard HMM approach can be improved by GMs with relaxed conditional independence statements
- More models available...

References

- Jeffrey A. Bilmes, *'Graphical Models and Automatic Speech Recognition'*, 2003
- Kevin Patrick Murphy, *'Dynamic Bayesian Networks: Representation, Inference and Learning'*, 2002
- Jeffrey A. Bilmes, *'Dynamic Bayesian Multinets'*, 2000
- Jeffrey A. Bilmes, *'Data-Driven Extensions to HMM Statistical Dependencies'*, 1998
- GMTK: The Graphical Models Toolkit, `http://ssli.ee.washington.edu/~bilmes/gmtk/`