

Infrastructural Goals in Speech & Language Systems

Jeff A. Bilmes

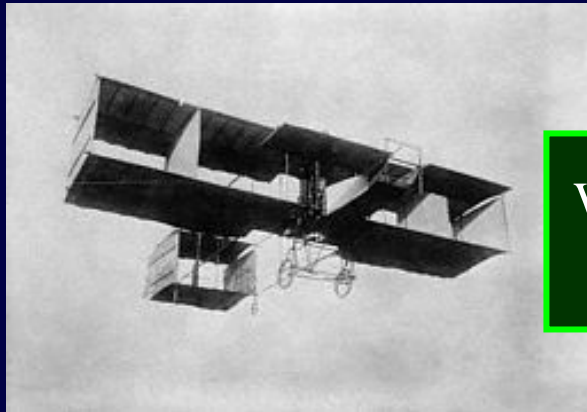
University of Washington

Department of EE, SSLI-Lab

Outline

- I. Ideal Infrastructure: Criteria
- II. A Candidate: Graphical Models for Speech/Language
- III. GMTK: An infrastructure for graphical model based speech and language processing

The State of Speech/Language Technology



We are about here.



Initial Technology

Useable Technology

Great Technology

Goal: Move to the right

Airplane analogy
by Hynek Hermansky

How to move to the right?

- Chin Lee's NSF Symposium on Next Generation ASR
- More research and more funding in speech/language processing
- Research of non-standard more risky methods
- Sustained long-term effort: to get over *the engineering bottleneck*
 - Towards increasing ASR error rate (Boulard, Hermansky, Morgan)
- New Infrastructure: *an enabling factor*

Why Have Infrastructure?

- No need to re-invent the wheel – can stand on each others shoulders

Infrastructure is not always good

- Increased tendency to keep trying only slightly different things (maintains status quo).

What Makes Good Infrastructure?

(top eleven countdown)

11. **Compatible** with existing infrastructure
10. **Efficient/Fast Execution**
9. **Open Source**
8. **Extensible** - Modular/Clean Source Code
7. **Pedagogical/Good for Students & Learning**
6. **Easy to Use/Learn by researcher**
5. **Alive** - maintained by someone
4. **Good Documentation**
3. **Free of Bugs**
2. **Allows Rapid Turnaround** (time from idea to solution)
1. **Generality** - widely varying techniques easily specifiable within one paradigm

Graphical Models

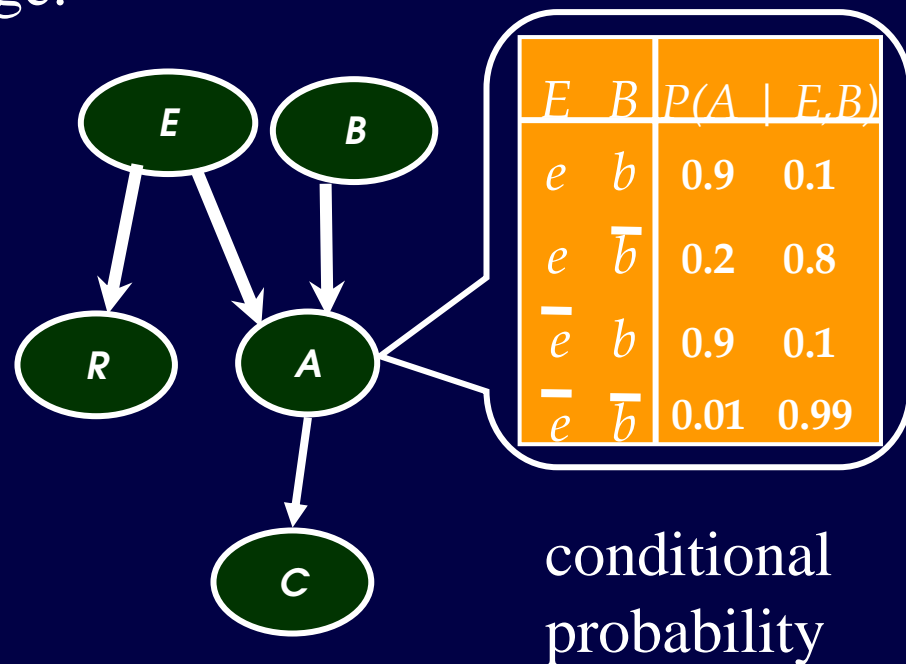
Better (more parsimonious) representations of joint probability distributions via the formal mathematical utilization of conditional independence via a visual language.

Graphs may be directed (DAG) or undirected

- Nodes - random variables
- Edges - direct influence
- Lack of edges – conditional independence

Graph defines a unique factorization of distribution

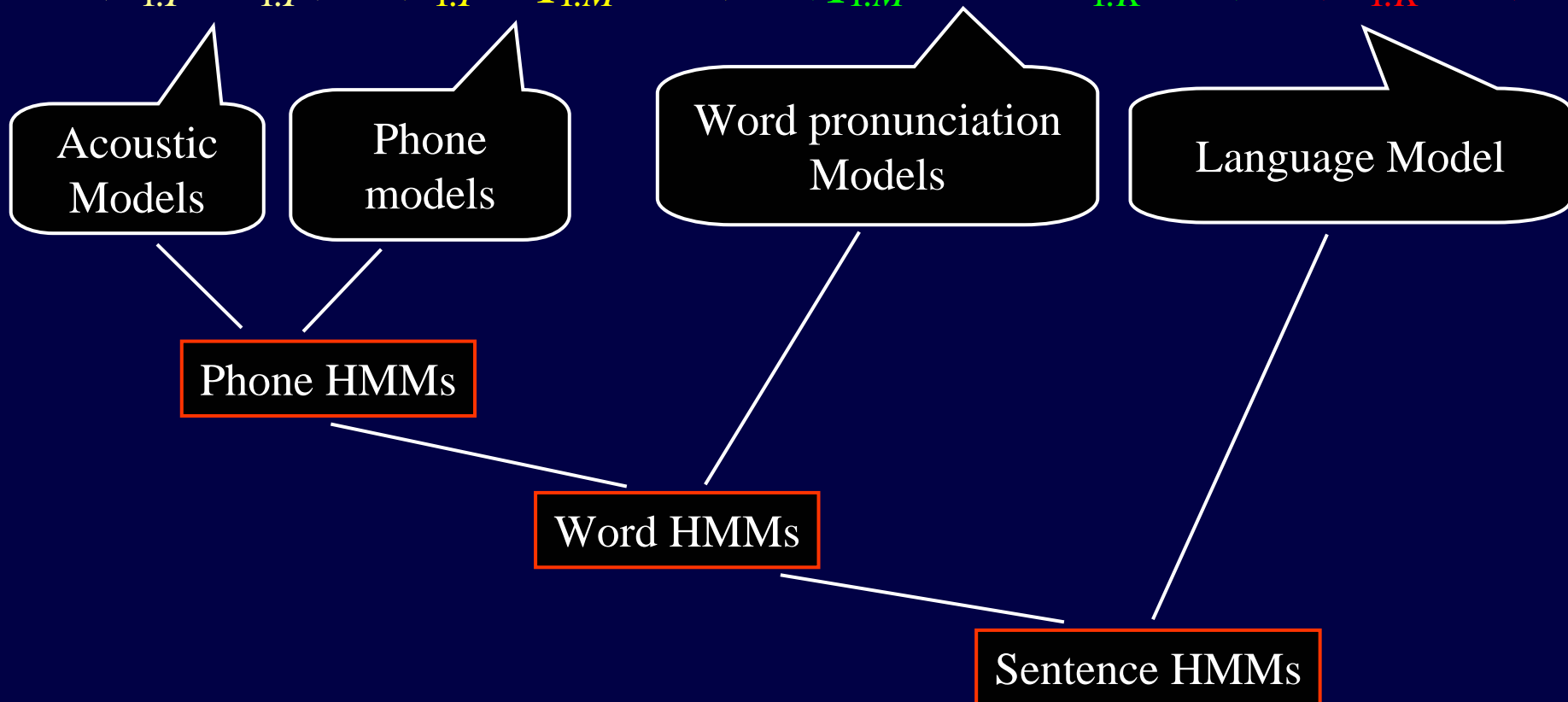
$$P(B, E, A, C, R) = P(B)P(E)P(A | B, E)P(R | E)P(C | A)$$



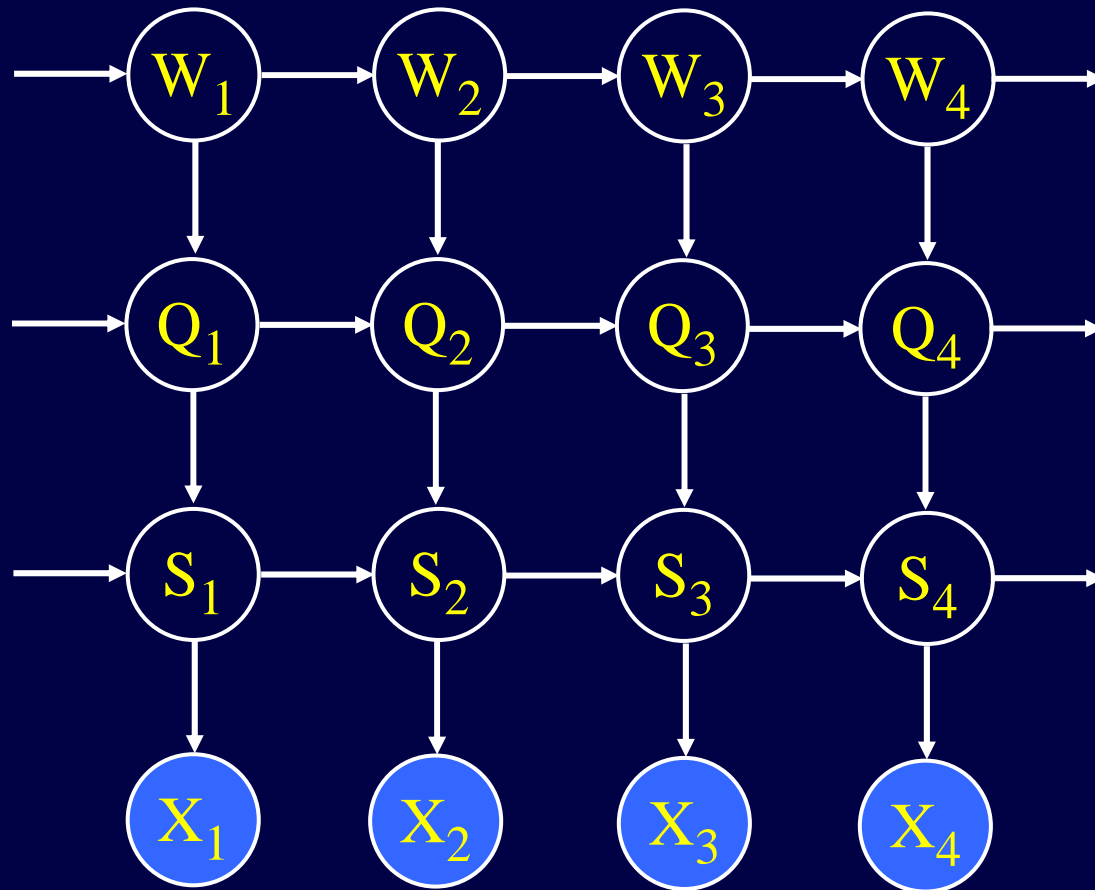
Generative Models of Speech

$$P(x_{1:T}, s_{1:T}, q_{1:M}, M, w_{1:K}, K) =$$

$$P(x_{1:T} | s_{1:T})P(s_{1:T} | q_{1:M}, M)P(q_{1:M}, M | w_{1:K}, K)P(w_{1:K}, K)$$

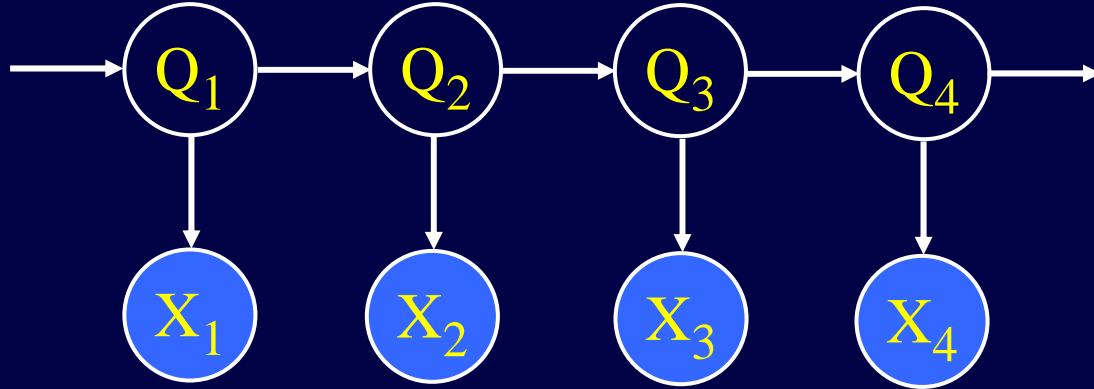


In other words, ASR has used Hierarchical GMs for years, ...



Hidden Markov Models

- ... but, in existing speech systems, all of this complexity gets **implicitly** wrapped up (flattened) into an HMM



- Number of flattened states is strongly dependent on language model
 - Bi-gram language model: $P(w_t | w_{t-1})$
 - Tri-gram language model: $P(w_t | w_{t-1}, w_{t-2})$

Why Graphical Models for Speech and Language Processing

- Expressive but concise way to describe properties of families of distributions
- **Rapid** movement from novel idea to implementation (with the right toolkit)
- GMs encompass many existing techniques used in speech and language processing but GM space is only barely covered
- Holds promise to improve upon the ubiquitous HMM, e.g., using factored representations.
- Dynamic Bayesian networks and dynamic graphical models can represent important structure in “natural” time signals such as speech/language.

Four Main Goals for GMs in Speech/Language

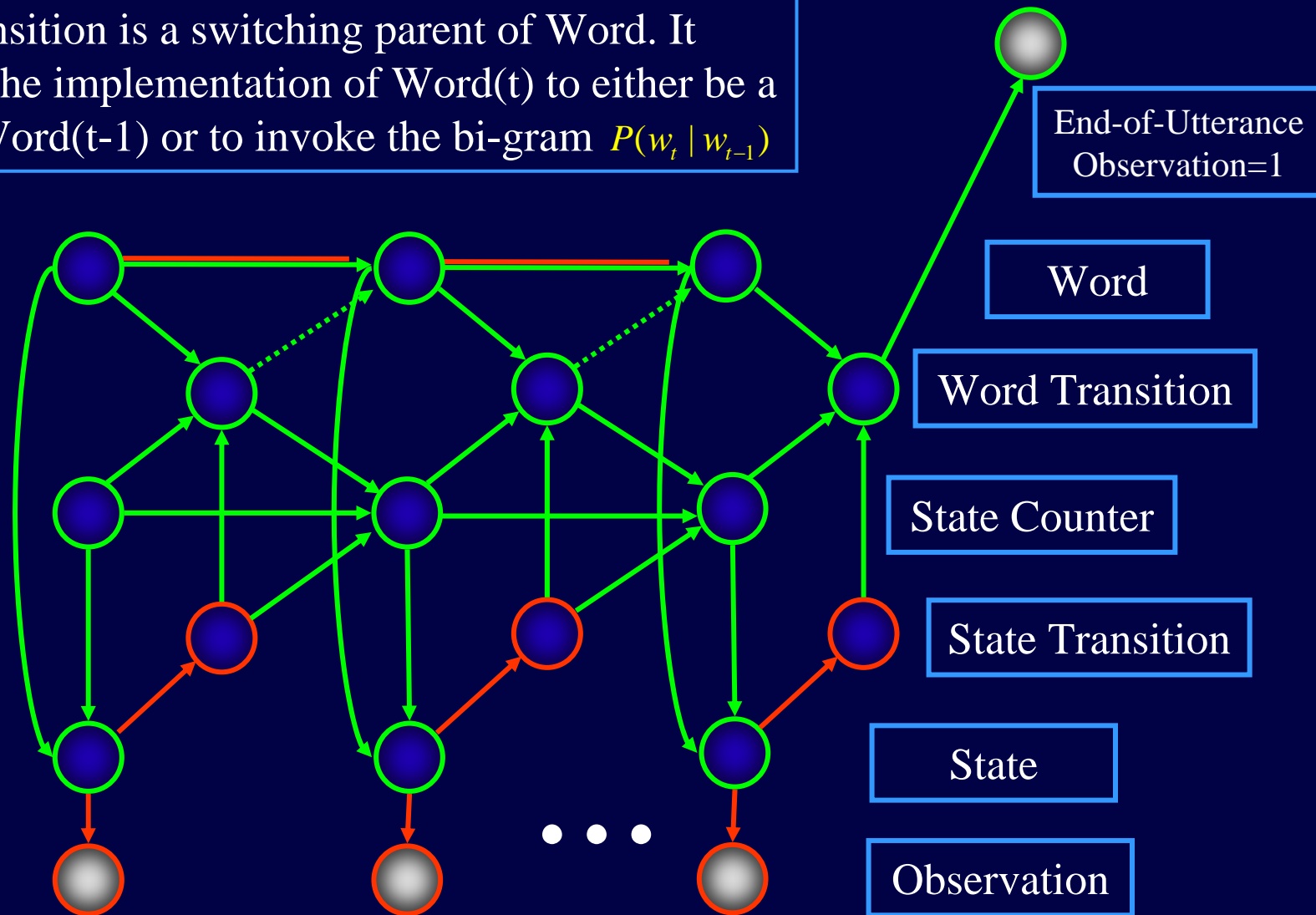
1. ***Explicit Control***: Derive graph structures that themselves *explicitly* represent control constructs
 - E.g., parameter tying/sharing, state sequencing, smoothing, mixing, backing off, etc.
2. ***Latent Modeling***: Use graphs to represent *latent information* in speech/language, not normally represented.
3. ***Observation Modeling***: represent structure over observations.
4. ***Structure learning***: Derive *structure* automatically, ideally to improve error rate while simultaneously minimizing computational cost.

Explicit bi-gram Decoder

WordTransition is a switching parent of Word. It switches the implementation of Word(t) to either be a copy of Word(t-1) or to invoke the bi-gram $P(w_t | w_{t-1})$

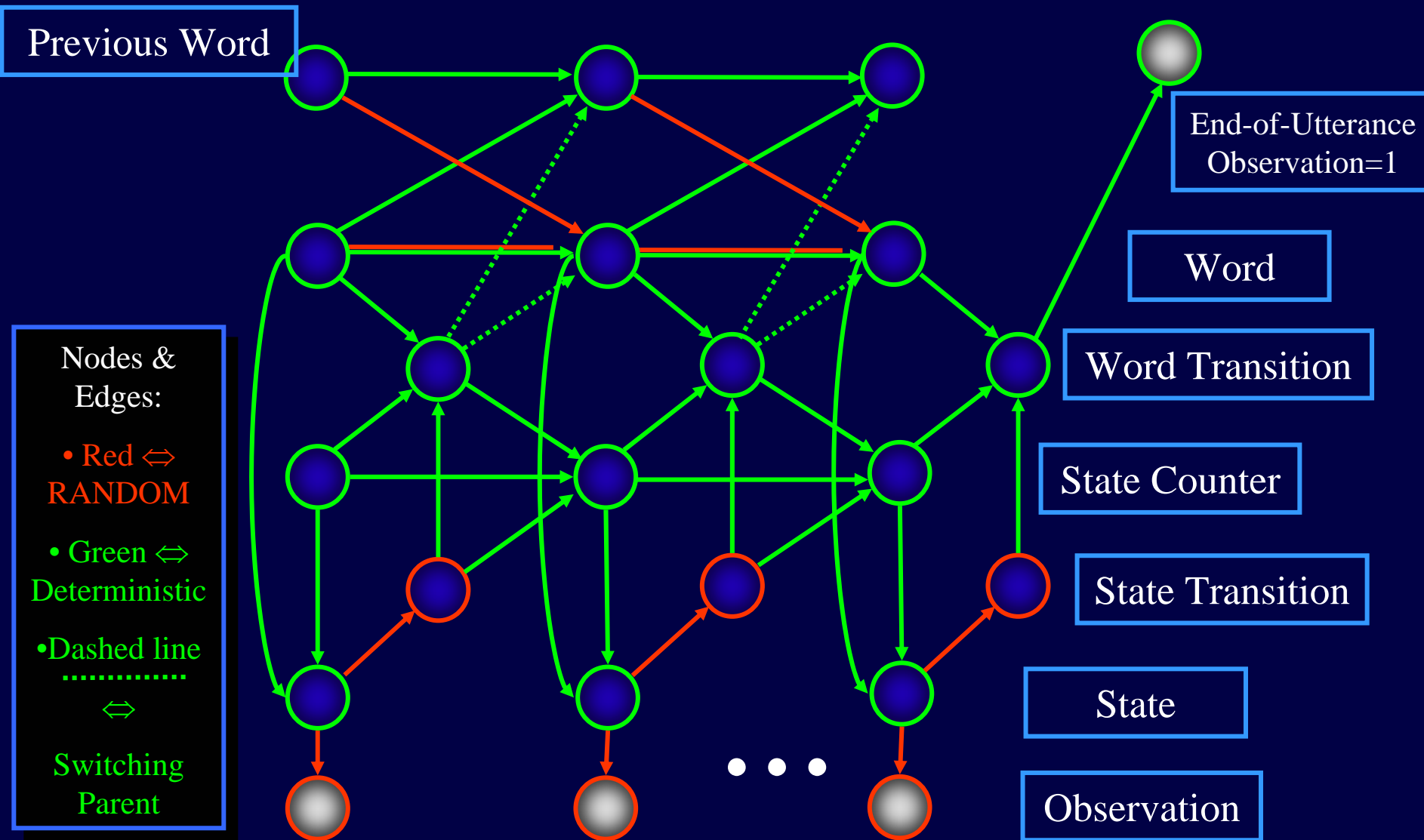
Nodes & Edges:

- Red \Leftrightarrow RANDOM
- Green \Leftrightarrow Deterministic
- Dashed line $\dots\dots\dots$
- \Leftrightarrow Switching Parent

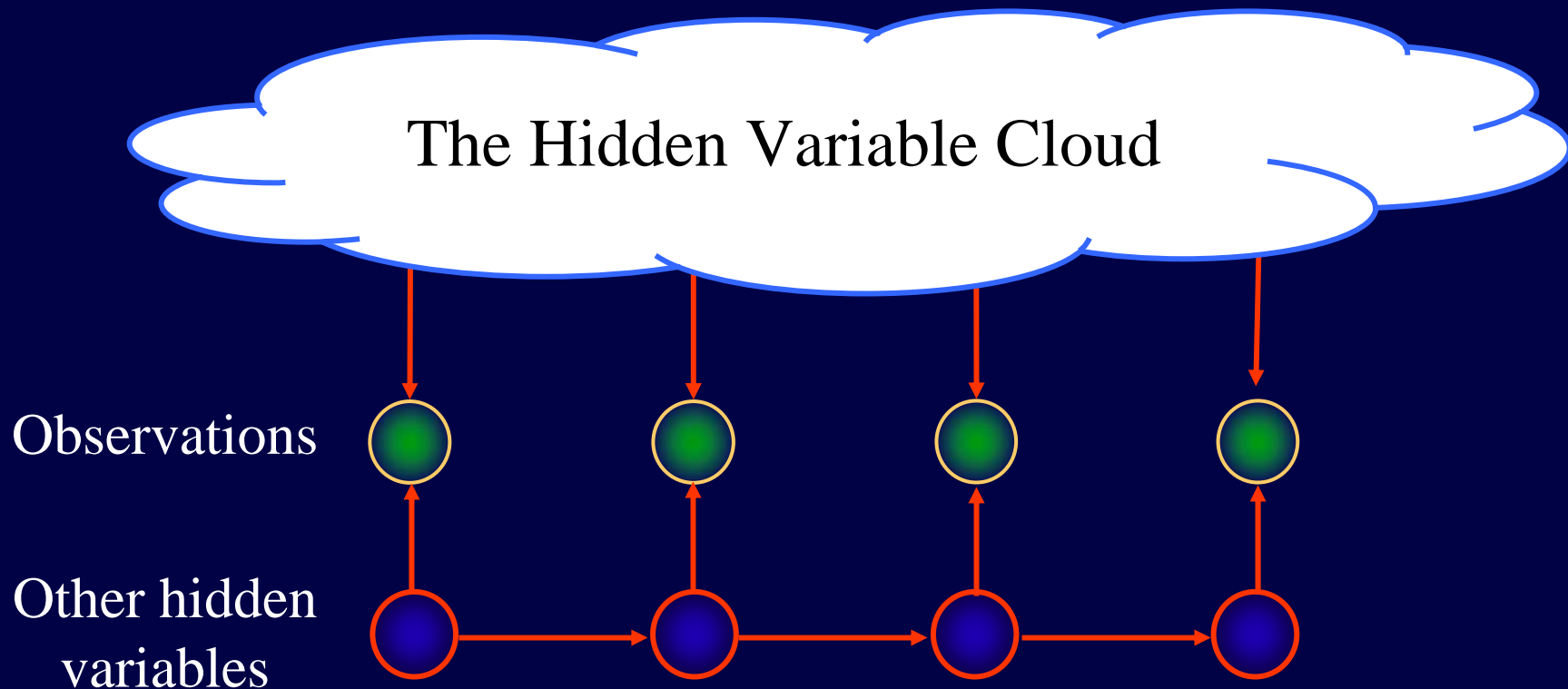


Explicit tri-gram Decoder

$$P(w_t | w_{t-1}, w_{t-2})$$

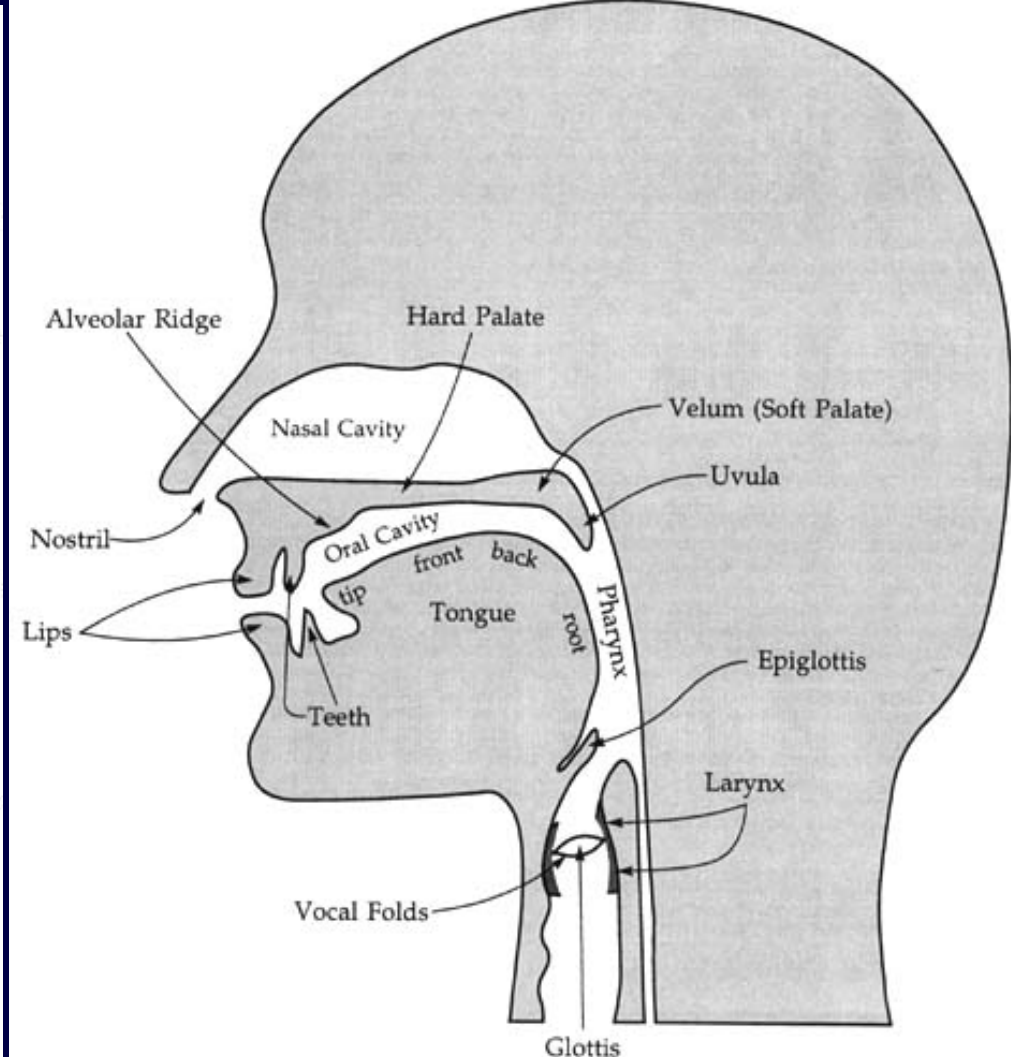
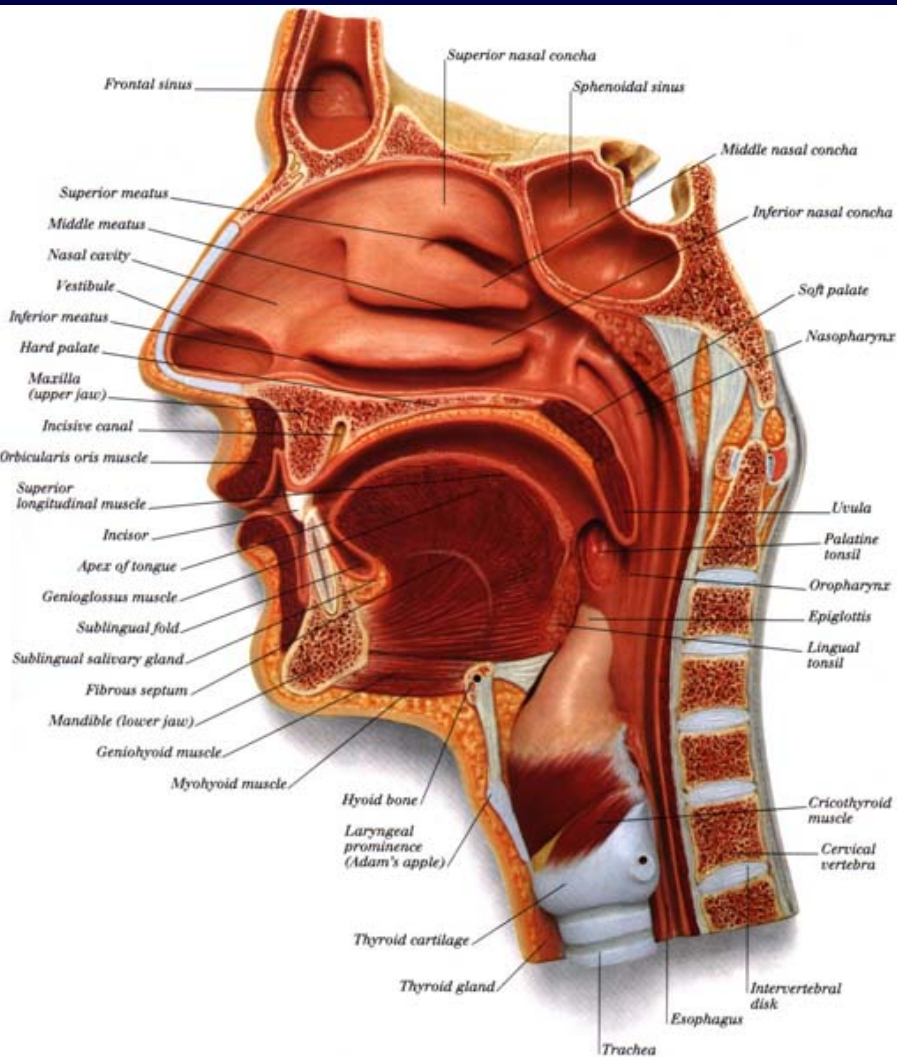


Latent X Modeling



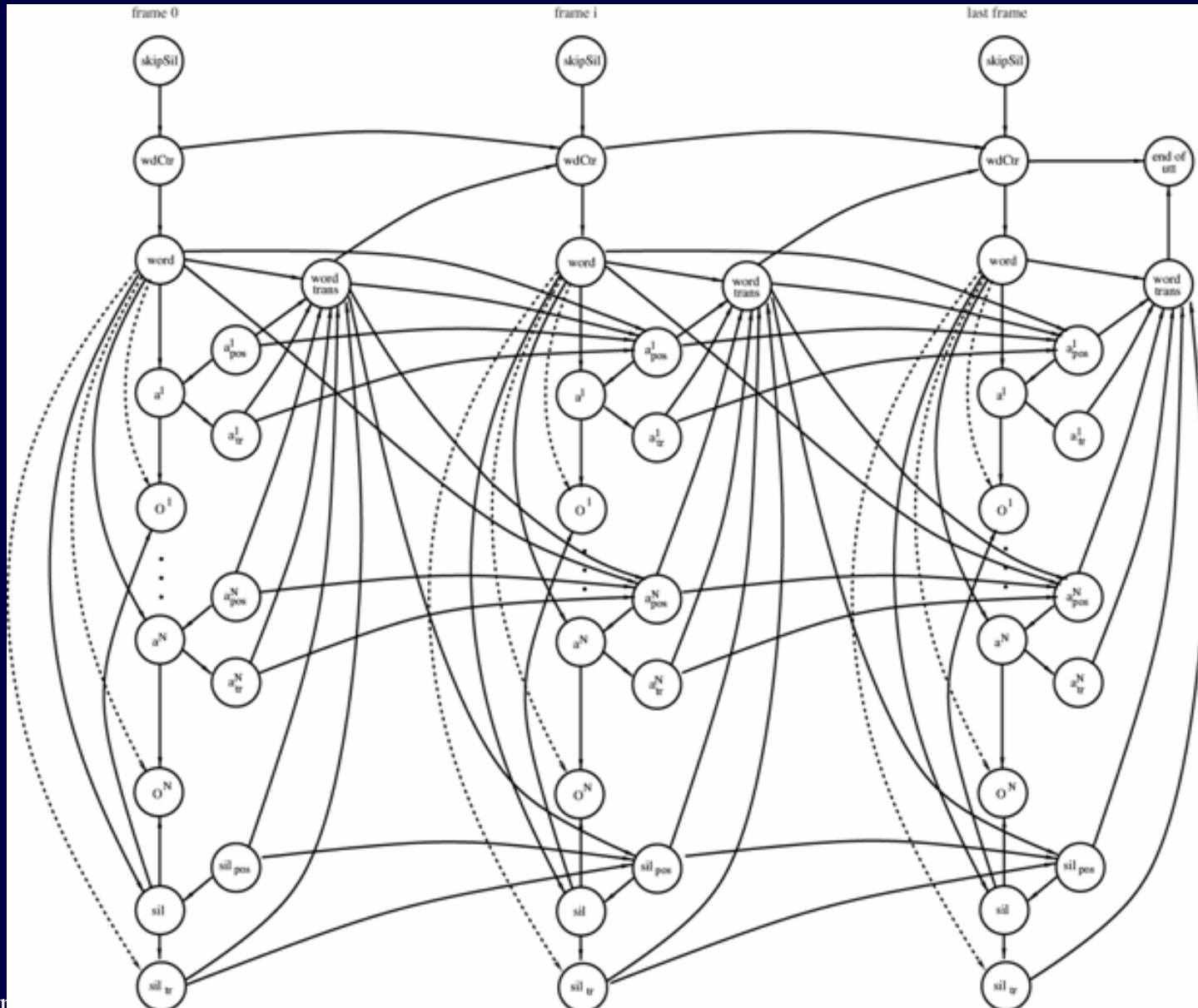
- Where X = gender, speaker cluster, speaking rate, noise condition, accent, dialect, pitch, formant frequencies, vocal tract length, etc.
- We elaborate upon latent articulatory modeling...

Ex: Latent Articulatory Modeling



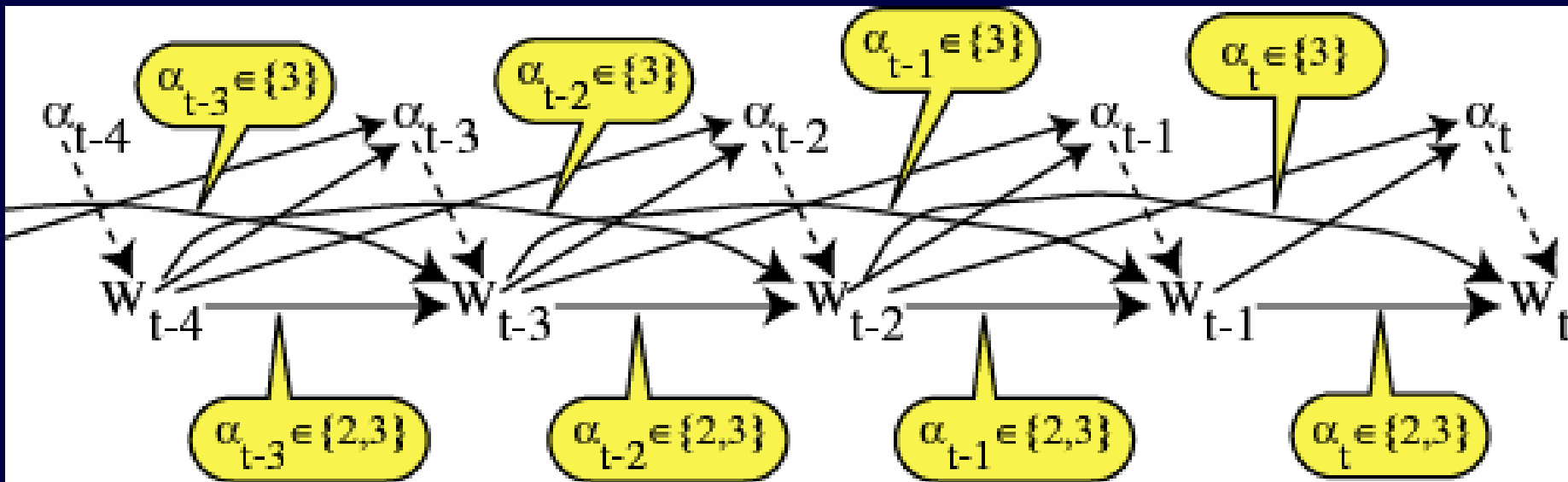
Pictures from Linguistics 001, University of Pennsylvania

Phone-free Articulatory Graph (by Karen Livescu)



Conditional mixture tri-gram

$$\begin{aligned} P(w_t | h_t) &= P(\alpha_t = 1 | w_{t-1}, w_{t-2}) P(w_t) \\ &+ P(\alpha_t = 2 | w_{t-1}, w_{t-2}) P(w_t | w_{t-1}) \\ &+ P(\alpha_t = 3 | w_{t-1}, w_{t-2}) p(w_t | w_{t-1}, w_{t-2}) \end{aligned}$$

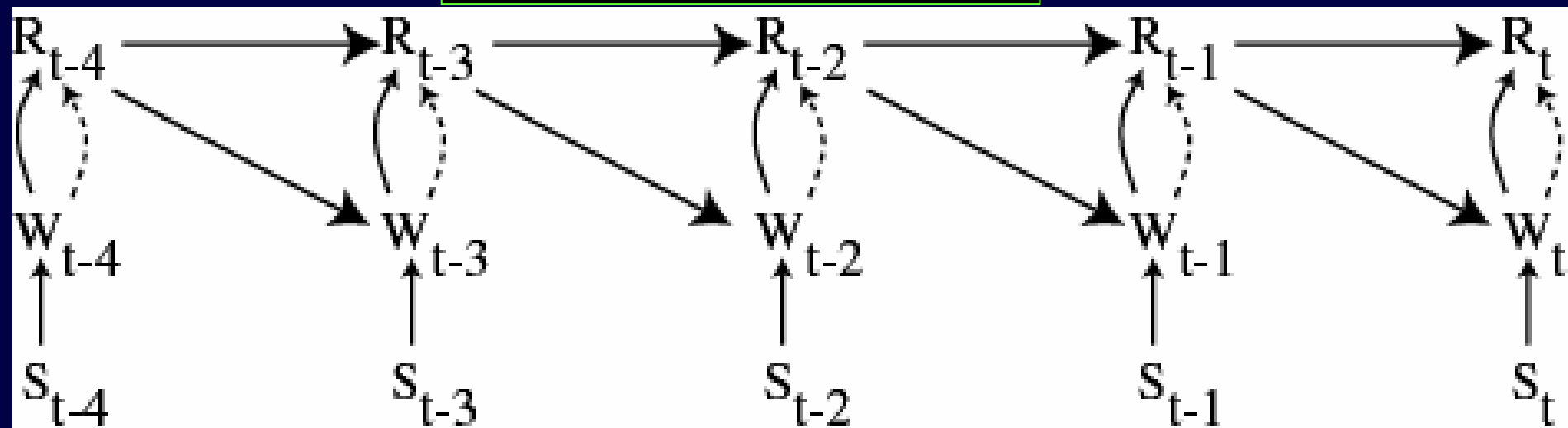


Skip Bi-gram

$$p(r_t | w_t, r_{t-1}) = \begin{cases} \delta_{r_t=r_{t-1}} & \text{if } w_t = \text{sil} \\ \delta_{r_t=w_t} & \text{if } w_t \neq \text{sil} \end{cases}$$

$$p(w_t | s_t, r_{t-1}) = \begin{cases} \delta_{w_t=\text{sil}} & \text{if } s_t = 1 \\ p_{\text{bigram}}(w_t | r_{t-1}) & \text{if } s_t = 0 \end{cases}$$

$$p(s_t = 1) = \Pr(\text{silence})$$

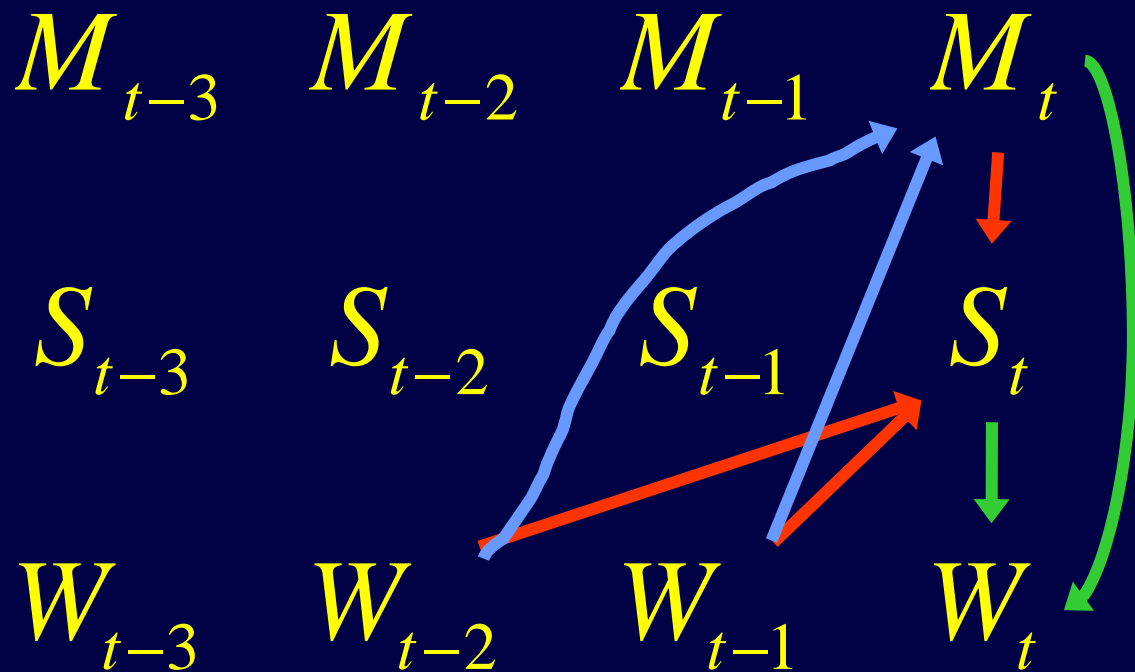


Factored Language Models

(work with Katrin Kirchhoff)

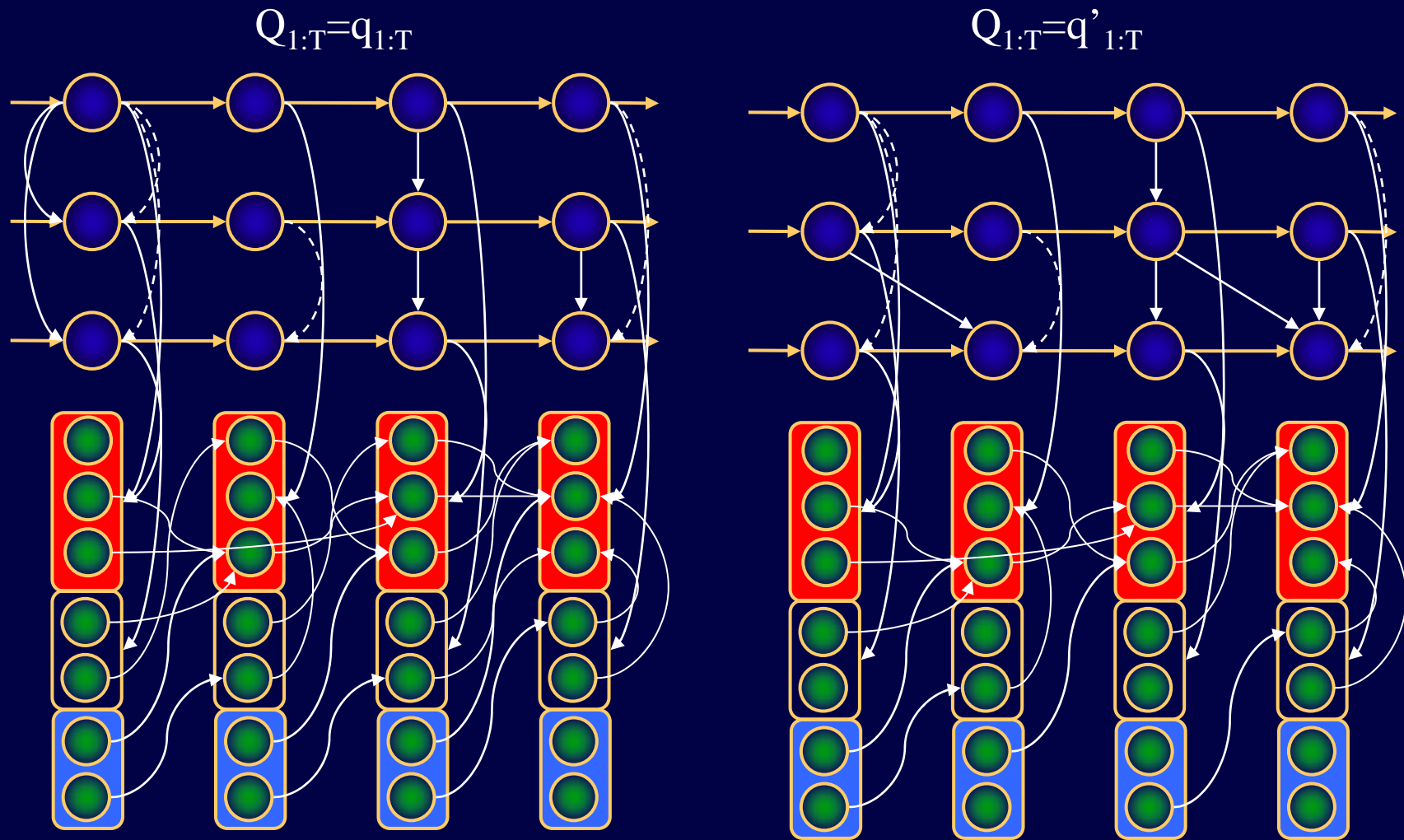
- Decompose words into smaller morphological or class-based units (e.g., morphological classes, stems, roots, patterns, or other automatically derived units).
- Produce probabilistic models over these units to attempt to improve language modeling accuracy and parameter estimation
- Clearly useful for highly inflected languages such as German, Arabic, Hebrew, etc.
- Potentially useful for inflectionally poor languages such as English.

Words, Stems (root+pattern), & Morphological classes



$$P(w_t | s_t, m_t) \quad P(s_t | m_t, w_{t-1}, w_{t-2}) \quad P(m_t | w_{t-1}, w_{t-2})$$

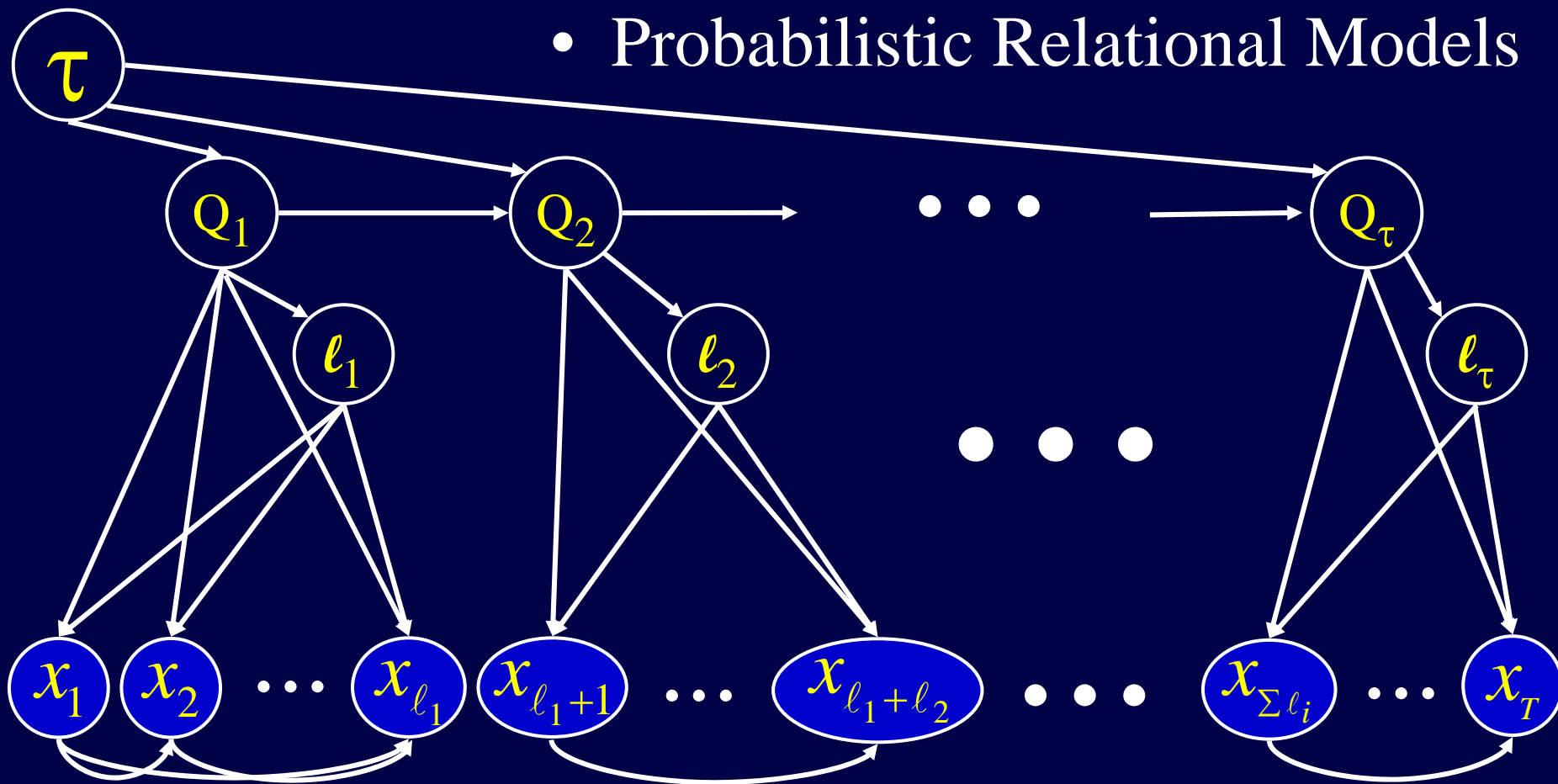
Multi-stream buried Markov models



Segment Models as GMs

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

- Probabilistic Relational Models



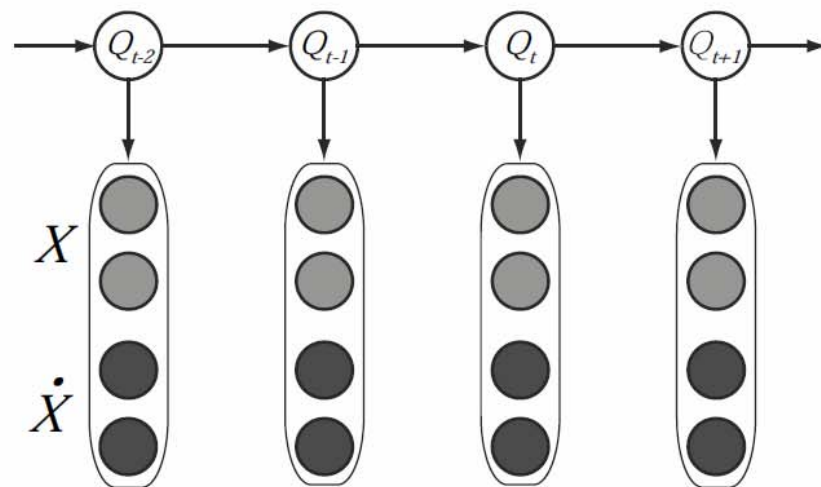
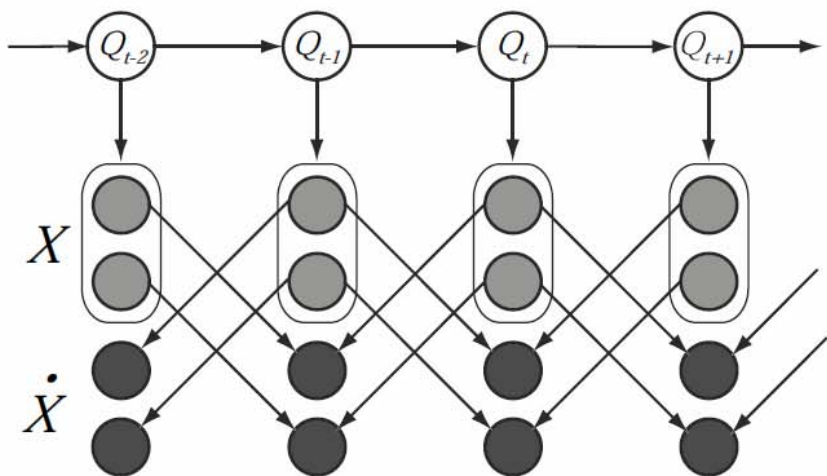
Discriminatively Structured Graphical Models

- **distinct** from (obviates??) discriminative parameter training
- Overall goal: **model parsimony**,
 - achieve same or better performance with same or fewer parameters.
- **represent only the “unique” dependencies** of each class,
- be incapable of not discriminating

Why Do Delta Features Work?

- Generative model of delta features
- deltas and hidden variables conditionally independent

- Wrong is Right!!!
- Incorrect generative model is more discriminative!!



GMTK: Graphical Models Toolkit

- A GM-based software system for speech, language, and time-series modeling
- One system – Many different underlying statistical models (more than an HMM)
- Complements rather than replaces other ASR and GM systems (e.g., HTK, AT&T, ISIP, BNT, BUGS, Hugin, etc.)
- Freely available, to be open-source

GMTK is infrastructure

- GMTK does not solve speech and language processing problems, but provides tools to help to simplify testing modeling, and does so in novel ways.
- The space of possible solutions is huge, and its exploration has only just begun.

GMTK Features

GMTK has many features that make it useful

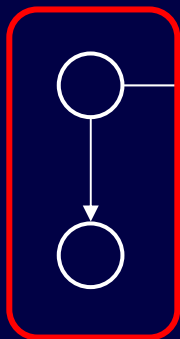
- Textual Graph Language
- Switching Parent Functionality
- Backwards time links
- Multi-rate models with extended DBN templates.
- Linear/Non-linear Dependencies on observations
- Arbitrary low-level parameter sharing (EM/GEM training)
- Gaussian Vanishing/Splitting algorithm.
- Decision-Tree-Based implementations of dependencies (deterministic and sparse)
- Full inference, single pass decoding possible on smaller tasks (current version)
- Sampling Methods
- Log space Exact Inference/Island algorithm – Memory $O(\log T)$

GMTK Structure file for HMM

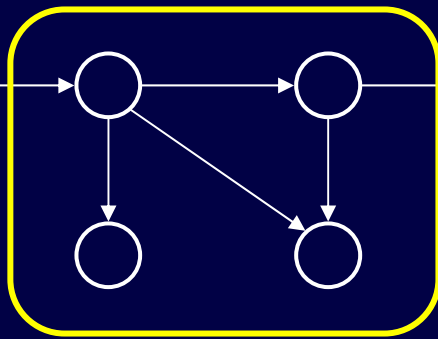
```
frame : 0 {
  variable : state {
    type : discrete hidden cardinality 4000;
    switchingparents : nil;
    conditionalparents : nil using DenseCPT("pi");
  }
  variable : observation {
    type : continuous observed 0:39;
    switchingparents : nil;
    conditionalparents : state(0) using mixGaussian mapping("state2obs");
  }
}
frame : 1 {
  variable : state {
    type : discrete hidden cardinality 4000;
    switchingparents : nil;
    conditionalparents : state(-1) using DenseCPT("transitions");
  }
  variable : observation {
    type : continuous observed 0:39;
    switchingparents : nil;
    conditionalparents : state(0) using mixGaussian mapping("state2obs");
  }
}
```

Multiframe Repeating Chunks

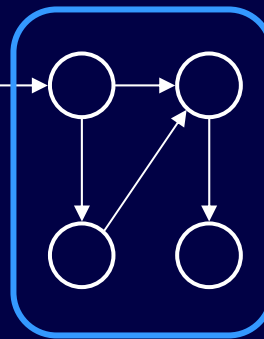
Prologue



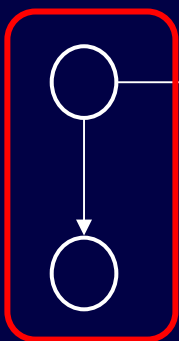
Repeating Chunk



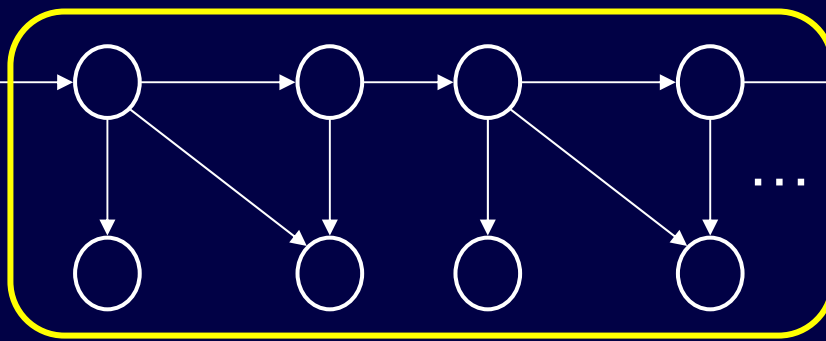
Epilogue



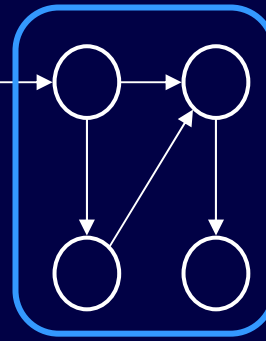
Prologue



Chunk Unrolled 1 time



Epilogue



Current GMTK Status

- I. System currently available (co-written with Geoff Zweig) at:
 - A. <http://ssli.ee.washington.edu/~bilmes/gmtk>
 - B. ~100 pages of documentation
 - C. Book chapter on use of graphical models for speech and language
 - D. JHU'2001 Workshop technical report
- II. New Re-Written Version (will be open source)
 - I. New Triangulation Engine running and ready (up to 400x faster) UAI'03
 - II. New GM data-structures (joins with tree-hash-lists, separator iterations, ancestral parent assignments, packed clique values)
 - III. Integration of DARPA and Factored Language Models
 - IV. Graphical Adaptation
 - V. Non-linear BMMs

The End

thank you!

- And thanks to the following people:
 - Alex Acero, Bill Byrne, Lee Deng, Chris Bartels, Ozgur Cetin, Karim Filali, Jim Glass, Steve Greenberg, Mary Harper, Hynek Hermansky, Mike Jordan, Sanjeev Khudanpur, Katrin Kirchhoff, Chin Lee, Karen Livescu, Brian Lucena, Nelson Morgan, Mari Ostendorf, Michael Picheny, Thomas Richardson, Geoff Zweig