Towards Auditory Scene Analysis Using Probabilistic Graphical Models

Taesu Kim, Hagai Attias, Te-Won Lee

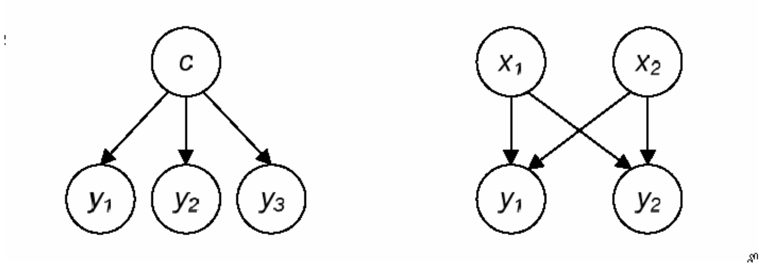
Motivation

- Find a Systematic and Principled Approach
- Learn from data
- Integration of Issues:
  - Signal separation and localization
  - Classification and clustering
  - Use of prior information and learned features
  - Use of additional sensing modalities (video, touch)

- Provide solutions to very complex scene analysis problems

Probabilistic Graphical Models

- Representation:
  - Nodes (variables), edges (dependencies)
  - Directed acyclic graphs (DAG)
  - Conditional probabilities
  - Joint probability

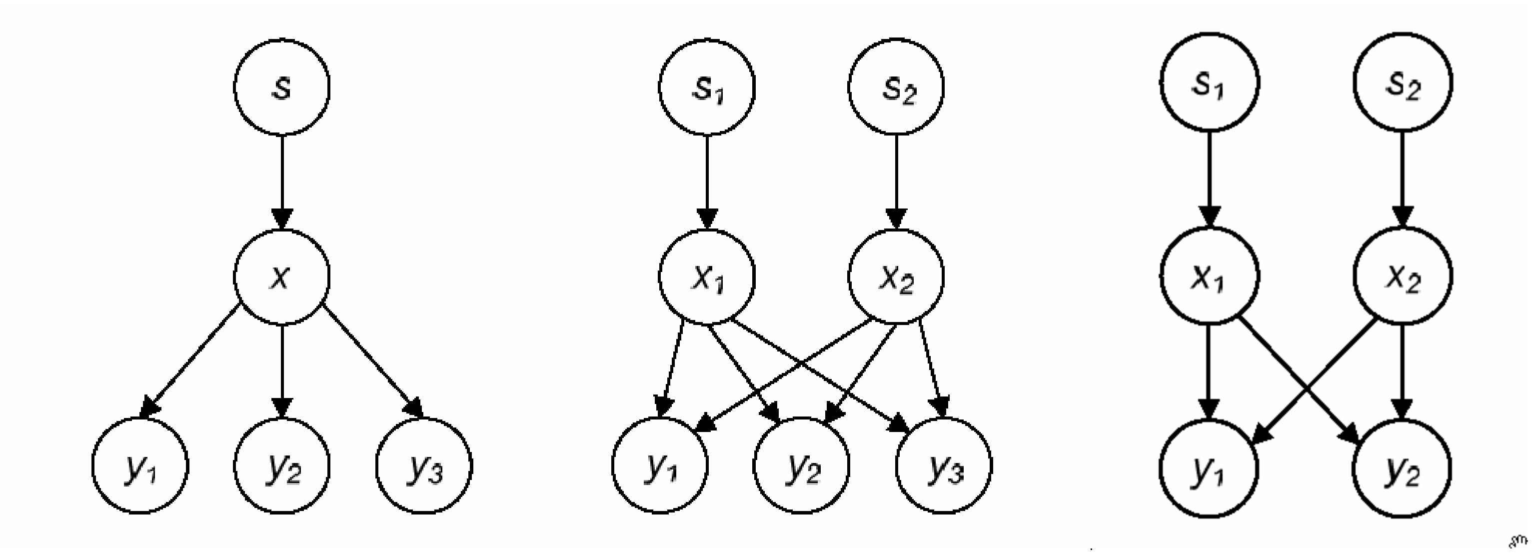


Supervised Classification   Unsupervised Learning (ICA)

- Inference
  - Posterior distribution
  - Bayes' Rule

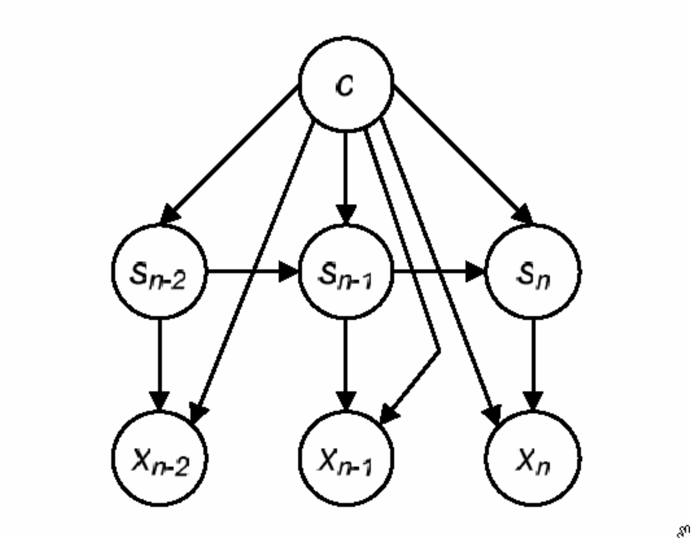
- Learning
  - Training model from data (estimating parameters from data)
  - Maximum likelihood
  - Expectation Maximization (EM) algorithm (iterative maximum likelihood)
    - E: Inference step: compute posterior distribution given current parameter values
    - M: Learning step: re-compute parameters using new posterior

Multiple Source – Multiple Observations

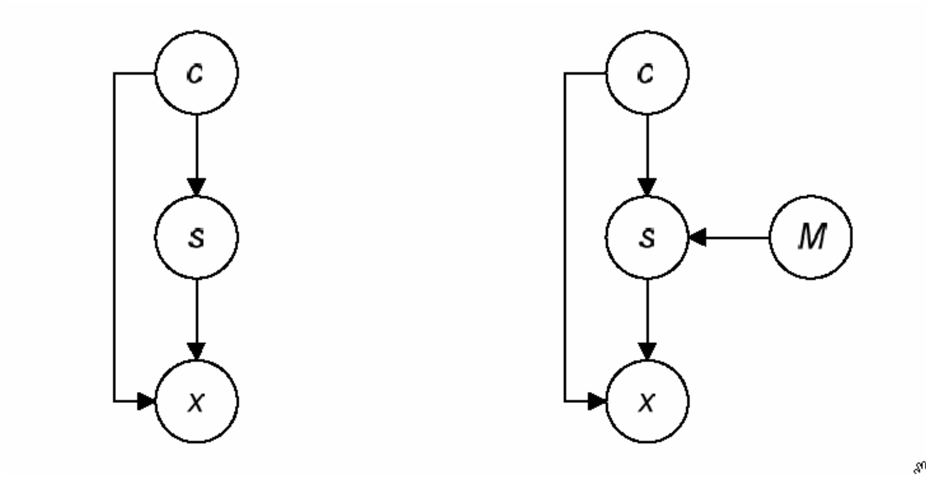


1) Single source, multiple sensors (deconvolution)
2) Two sources, three sensors (undercomplete)
3) Two sources, two sensors (complete)
4) Overcomplete?
5) Single channel?

Signal Representation & Classification

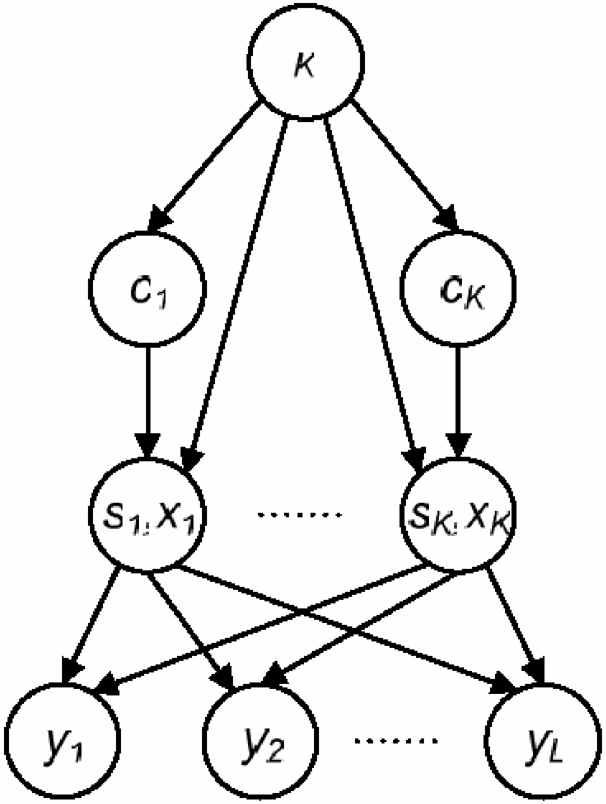


Source signals in subbands with corresponding class



(right) compact representation
(left) source, class, number of components

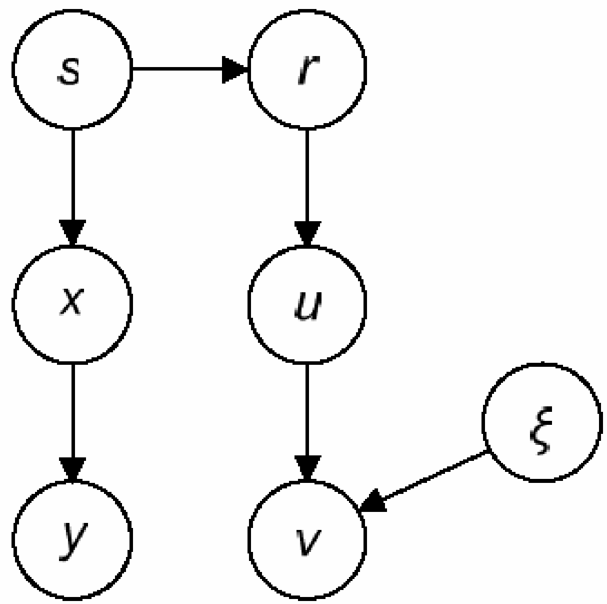
Classification of Auditory Signals



Add:

1) Model includes K sources
2) Separately drawn from a class

Learning with Audio-Visual Sensors



1) Microphone signal y from speech source (s,x)
2) Video signal v from face image u (shift & rotate),
   r is subspace variable depending on speech states s

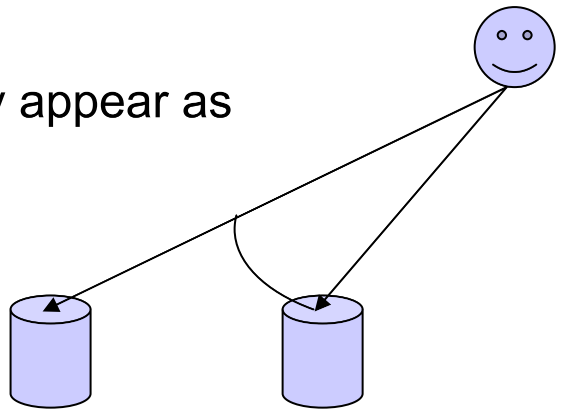
Models for Solving Complex Tasks

- Audio-Visual Scene Understanding:
  - Object and sources separation
  - Tracking, Localization, Identification

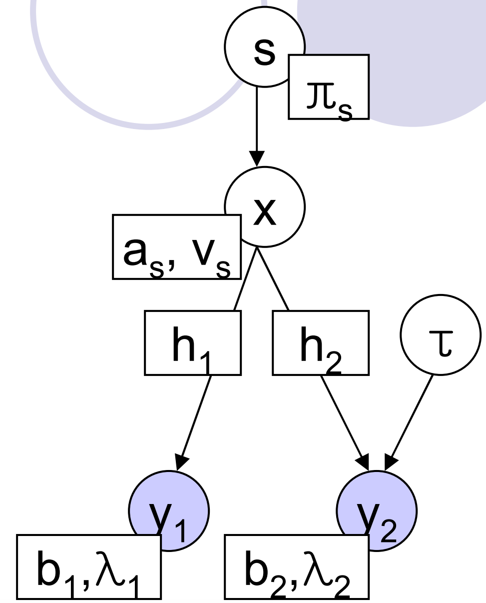
- Representation
  - Graphical models, temporal and spatial dependencies
- Learning
  - Approximate EM algorithms
- Inference
  - Variational Bayesian methods
- Evaluation
  - Database? Real time processing?

Probabilistic Graphical Model for Speaker Localization

- Motivation: Speech enhancement, video conferencing
- Key is robust time delay estimation in noisy environment
- Current approaches based on crosscorrelation among microphones
- Our approach:
  - Model joint distribution of the observed signals in terms of the unobserved signal originating from the speaker, distorted by unknown linear filtering due to reverberation as it propagated, contaminated by additive noise, and reaches the microphone with an unknown relative time delay.
- Use Probabilistic model of speech signals and noise
- Reverberation filter coefficients and relative time delay appear as parameters.



Microphone signal model



$$
\begin{aligned}
y_{1n} &= h_{1n} \star x_n + u_{1n} \\
y_{2n} &= h_{2,n-\tau} \star x_{n-\tau} + u_{2n}
\end{aligned}
$$

$$
b_{1n} \star u_{1n} = v_{1n} \ , \qquad b_{2n} \star u_{2n} = v_{2n}
$$

$$
\begin{aligned}
p(y_1 \mid x) &= \prod_n \mathcal{N}(b_{1n} \star y_{1n} \mid b_{1n} \star h_{1n} \star x_n, \lambda_1) \\
p(y_2 \mid x, \tau) &= \prod_n \mathcal{N}(b_{2n} \star y_{2n} \mid b_{2,n-\tau} \star h_{2,n-\tau} \star x_{n-\tau}, \lambda_2)
\end{aligned}
$$

- Noise model training : the noise parameters are estimated directly from pure noise segments obtained from silent parts of the microphone data.

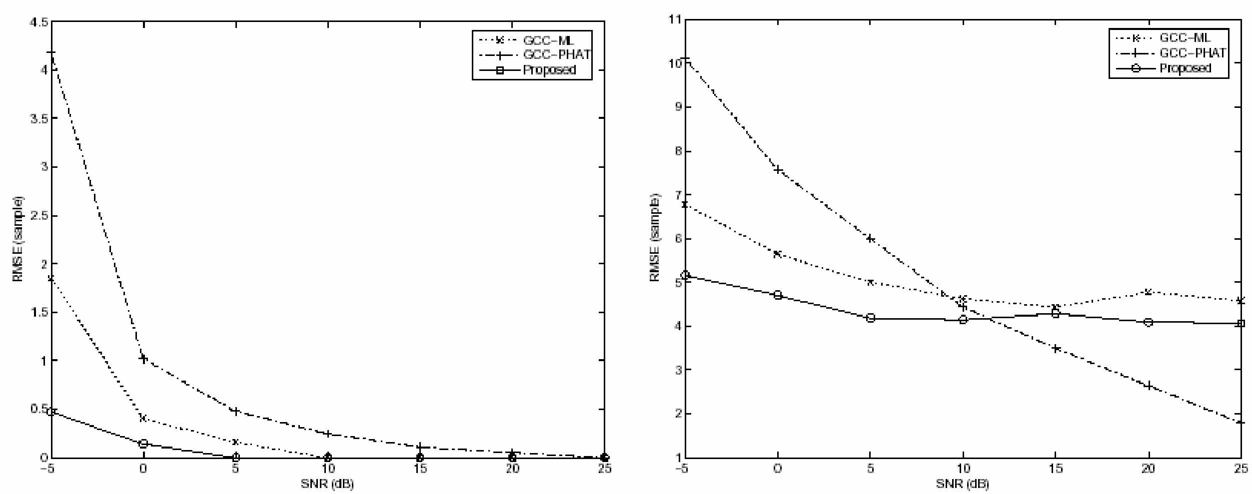
Experiments

- Performance measure

$$\text{bias} = |\tau - E[\hat{\tau}]| \,, \qquad \text{variance} = \text{Var}[\hat{\tau}]$$

$$\text{RMSE} = \sqrt{\text{bias}^2 + \text{variance}}$$

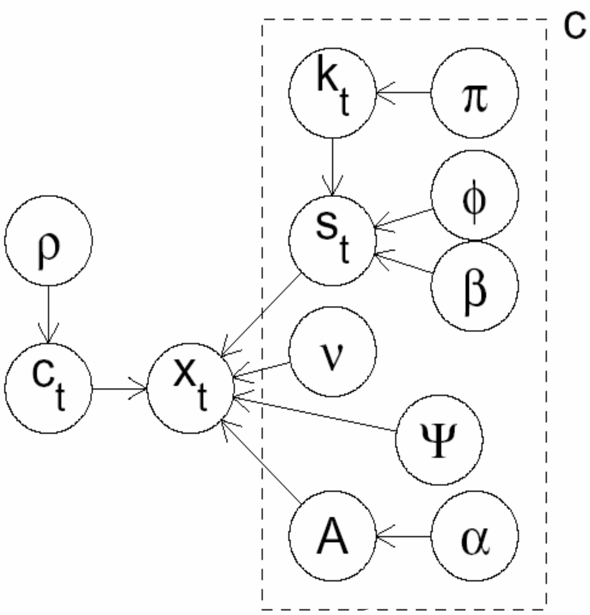
- Results : simulated data



(a) Reverberation time 0ms (b) Reverberation time 150ms

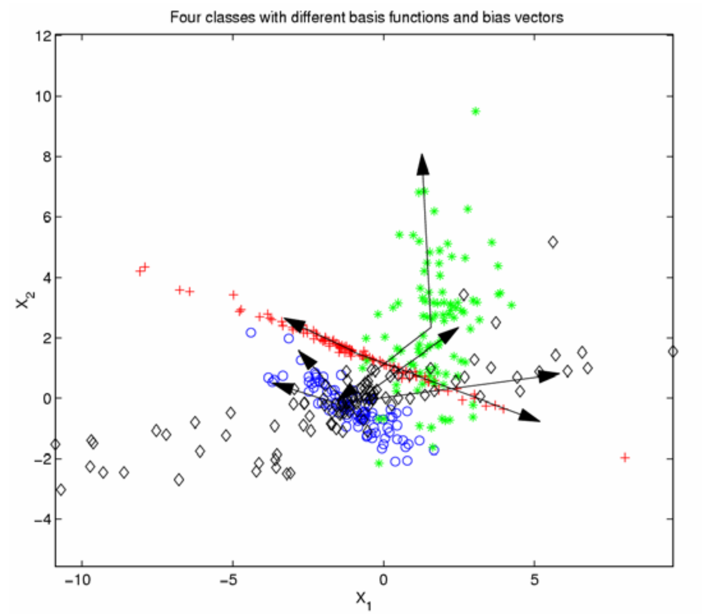
Single Channel Signal Separation

Graphical representation of generative model



x: Observed variable
k, s, c: Hidden variables
Rest: model parameters

x: measured signal
A: sub signal dictionary
s: activation coefficients
c: class of pattern set
k: number of patterns



Chan, Lee, Sejnowski; 2002

Conclusions and Future Work

- Probabilistic graphical models may provide systematic and principled approach to scene analysis
- Extension of source localization model to multiple sources, overcomplete representation
- To be robust to non-stationary conditions, extend to track filters and noise parameters as they change in time
- Learn correlated information from audio visual information