香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Sparse Models for Speech Recognition

Weibin Zhang and Pascale Fung

Human Language Technology Center
Hong Kong University of Science and Technology

Outline

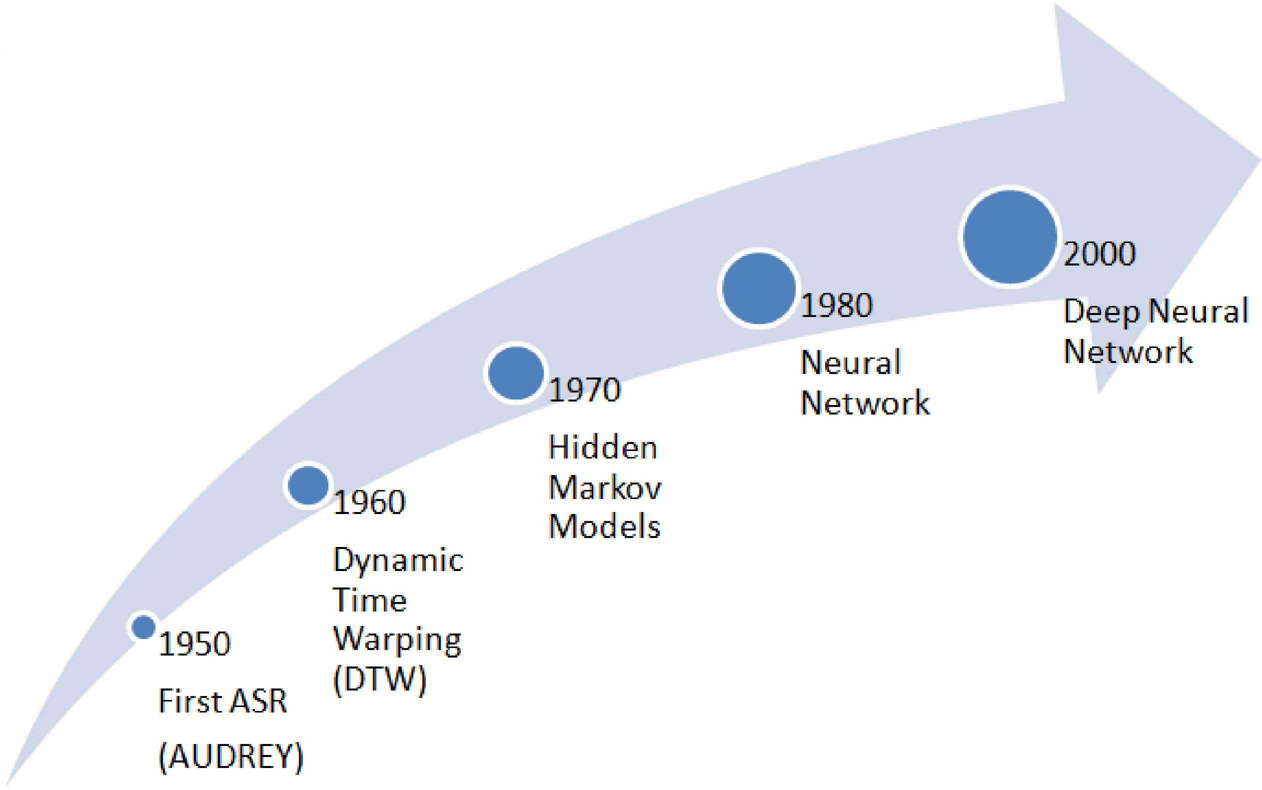
- Introduction to speech recognition
- Motivations for sparse models
- Maximum likelihood training of sparse models
- ML training of sparse banded models
- Discriminative training of sparse models
- Conclusions

Speech Recognition & its Applications

1. Automatic Speech Recognition (ASR):
   - Convert speech wave into text automatically
2. Applications:
   - Office/business systems:
   - Manufacturing
   - Telecommunications
   - Mobile telephony
   - Home Automation
   - Navigation
   - ………

History of ASR

- Technical Point of View



ASR Research -- Overview

- Statistical approaches lead in all area.
- Still big gap between human and machine performance...however
- Useful systems have been built which are changing the way we interact with the world

...within five years people will discard their keyboards and interact with computers using touch-screens and voice controls...

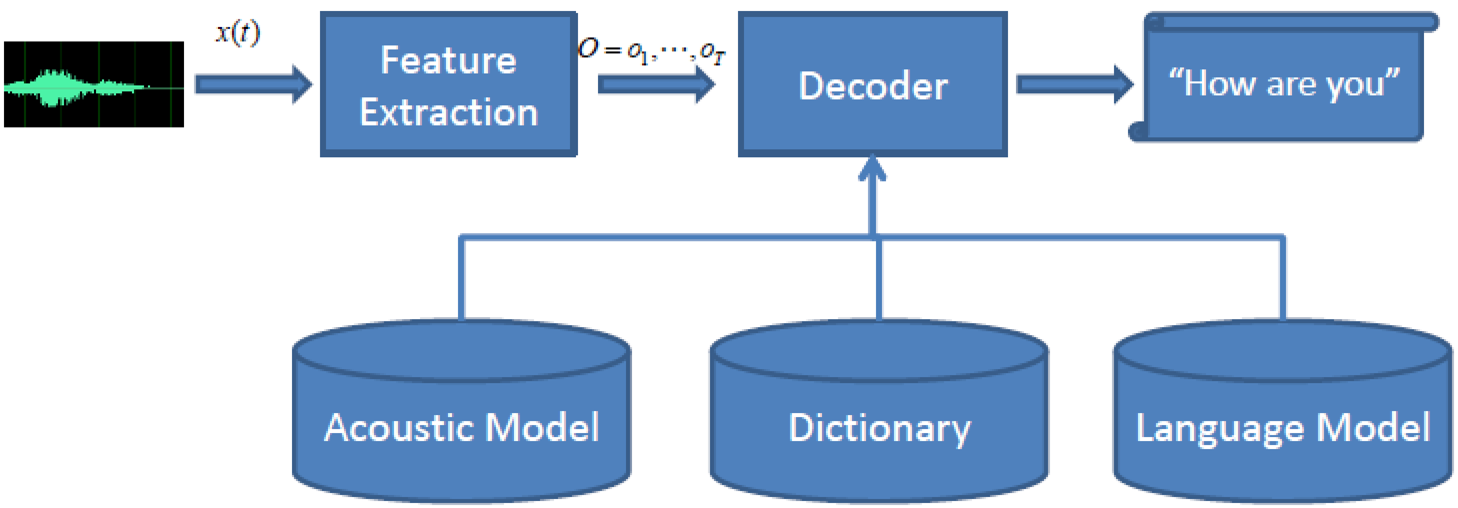
Bill Gates, Feb 2008

Statistical speech recognition system

Acoustic Model

$$\hat{W} = \arg\max_{W} P(W \mid O) = \arg\max_{W} \frac{P(O \mid W)P(W)}{P(O)} = \arg\max_{W} P(O \mid W)P(W)$$

Language Model



Statistical speech recognition system

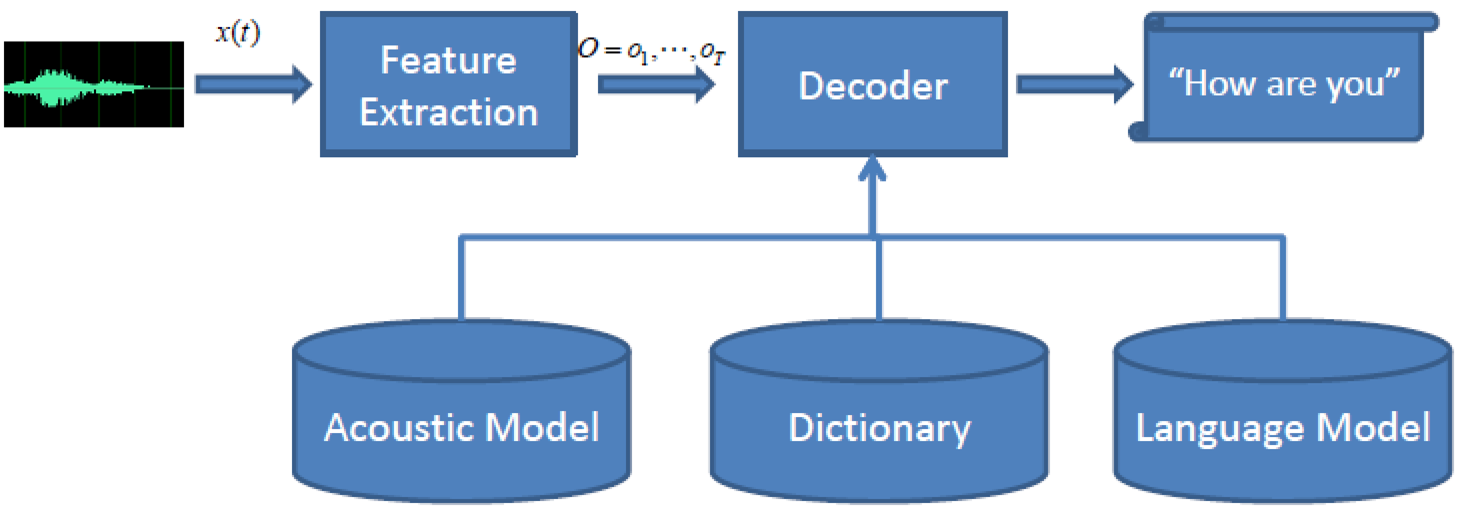
- Language Model:
  - P("recognize speech") >> P("wreck a nice beach")
- Dictionary:
  - Wreck r e k
  - Beach b i th
- Acoustic Model:
  - P(O|"recognize speech")

Statistical speech recognition system

$$\hat{W} = \arg\max_{W} P(W \mid O) = \arg\max_{W} \frac{P(O \mid W)P(W)}{P(O)} = \arg\max_{W} P(O \mid W)P(W)$$

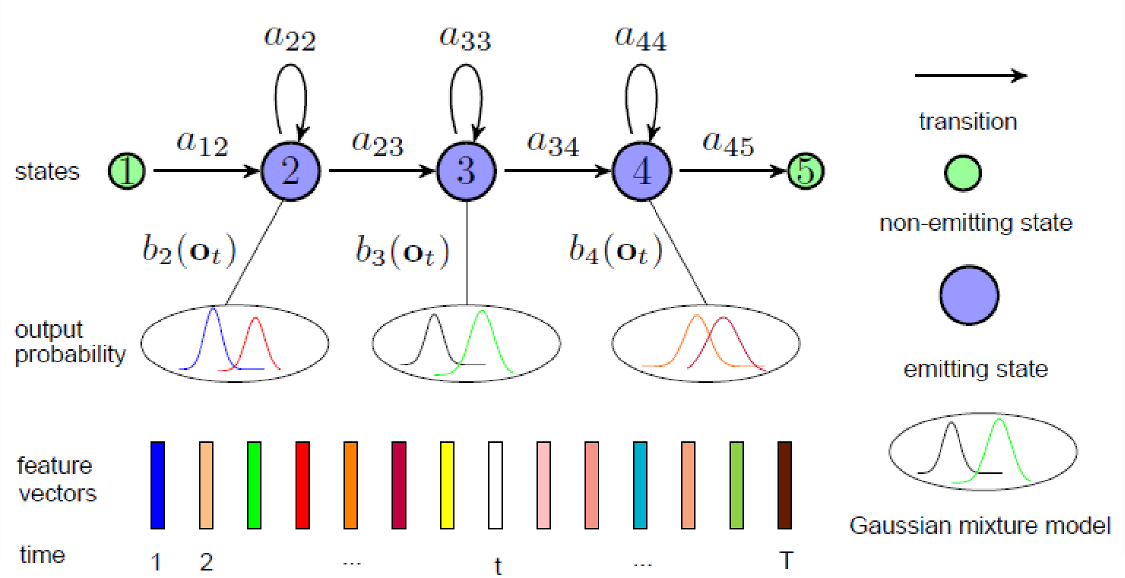
Acoustic Model — $P(O \mid W)$

Language Model — $P(W)$



Acoustic modeling

- *Left-to-right* hidden Markov models (HMMs)
- *GMM-HMM* based acoustic models
- $p(\boldsymbol{o}_t | s_j) = \sum_m c_{jm} N(\boldsymbol{o}_t : \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})$
- $\Theta = \{a_{ij}, b_j(\boldsymbol{o}_t)\} = \{a_{ij}, c_{jm}, \boldsymbol{u}_{jm}, \boldsymbol{\Sigma}_{jm}\}$



Evaluation of ASR system

- Word error rate (WER) = 1 – accuracy

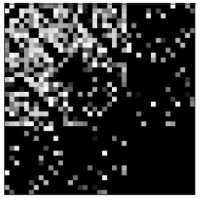
$$WER = \frac{S + D + I}{N}.$$

- Real time factor (RTF)

$$RTF = \frac{\text{decoding time}}{\text{duration of the utterance}}.$$

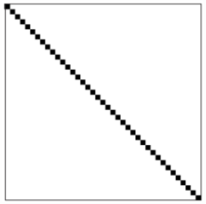
Covariance modeling

Full covariance matrices



- Better if data is sufficient
- More computation
- Easily over fit

Diagonal covariance matrices



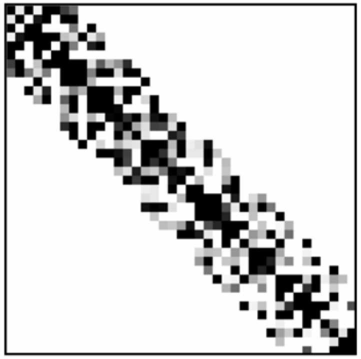
- Simple
- Features are independent
- More Gaussian components

Covariance modeling

Sparse banded inverse covariance matrices (*sparse models*)

- Alleviate over-fitting — Less Training data
- Less computation — $\frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$
- Reasonable model assumption (decorrelated features)

parameter type	number of parameters	percentage
transitions	1100	~ 0
weights	5686	0.2
means	221,754	4.7
precision matrices	4,435,080	95.1
total	4,663,620	100



ML training of sparse models

- Maximum likelihood (ML) training

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\{\log(P(\boldsymbol{O}|\boldsymbol{\Theta}))\}$$

- The proposed new objective function

$$L(\boldsymbol{\Theta}) = \log(P(\boldsymbol{O}|\boldsymbol{\Theta})) - \sum_{i=2}^{S-1}\sum_{m=1}^{M}\rho||\boldsymbol{C}_{im}||_1$$

- Auxiliary function:

$$Q(\boldsymbol{\Theta};\boldsymbol{\Theta}') = \sum_q\sum_m P(\boldsymbol{q},\boldsymbol{m}|\boldsymbol{\Theta}',\boldsymbol{O})\log(P(\boldsymbol{q},\boldsymbol{m},\boldsymbol{O}|\boldsymbol{\Theta})) - \sum_{i=2}^{S-1}\sum_{m=1}^{M}\rho||\boldsymbol{C}_{im}||_1$$

- Properties of the auxiliary function:
  - $L(\boldsymbol{\Theta}) - L(\boldsymbol{\Theta}') \geq Q(\boldsymbol{\Theta};\boldsymbol{\Theta}') - Q(\boldsymbol{\Theta}';\boldsymbol{\Theta}')$

Maximizing the auxiliary function

$$\max_{\Theta} Q(\Theta;\Theta')$$

$$P(\boldsymbol{q},\boldsymbol{m},\boldsymbol{O}|\Theta) = \prod_{t=1}^{T} a_{q_t q_{t+1}} c_{q_t m_t} b_{q_t m_t}(\boldsymbol{o}_t)$$

Forward and *backward* probabilities

Conditional independent assumptions of HMM

- The precision matrices can be updated using

$$\widehat{\boldsymbol{C}}_{im} = \underset{\boldsymbol{C}_{im} \succ 0}{\operatorname{argmax}}\{\operatorname{logdet}\boldsymbol{C}_{im} - \operatorname{trace}(\boldsymbol{S}_{im}\boldsymbol{C}_{im}) - \lambda||\boldsymbol{C}_{im}||_1\}$$

  - $\lambda = \frac{2\rho}{\gamma_{im}}$ and $\boldsymbol{S}_{im}$ is the sample covariance matrix.
  - Convex optimization or other more efficient methods (e.g. graphical lasso)

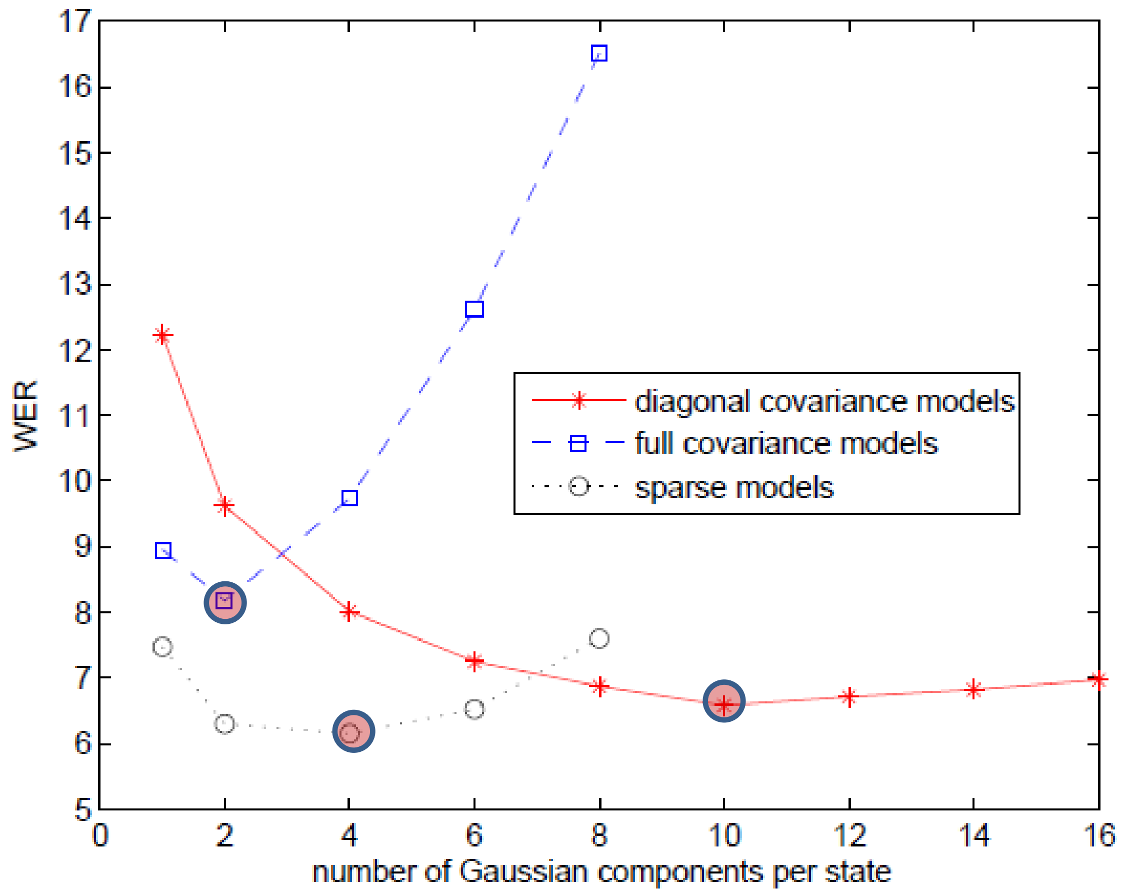
Experiments on the WSJ data

- Experimental setup
  - Training, development and testing data sets

data set	#speakers	#utterances	hours	vocab size
train(si84)	83	7134	14.5	8914
dev(Nov'92)	10	205	0.67	1270
eval(Nov'93)	8	330	0.41	988

  - Standard bigram language model
  - Feature vector: 39-dimension MFCC
  - 39 phonemes for English (39^3 triphones)
  - 2843 tied HMM states

Tuning results on the dev. data set

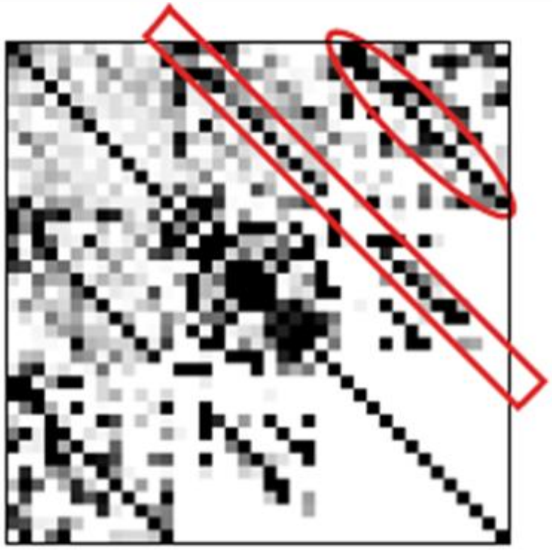


WER on the testing data set

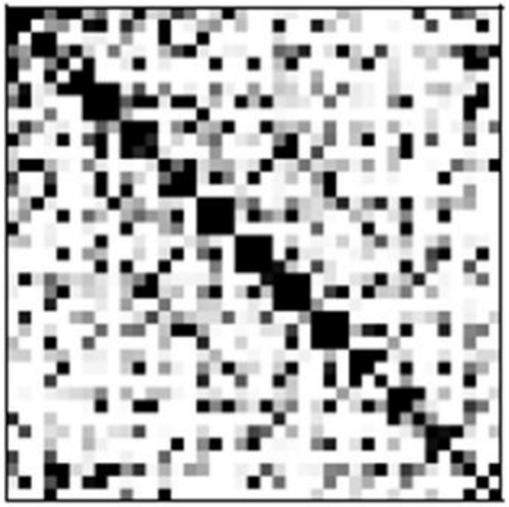
- Our result of 8.77% WER is comparable to the 8.6% WER reported in (Ko & Mak, 2011) using a similar testing configuration, but using 70 hours of training data

Model type	#Gaussians	WER	Rel. improv.	Significant?
Full	2	10.5	-7.1	No
Diagonal	10	9.84	----	----
Sparse	4	**8.77**	10.9%	Yes

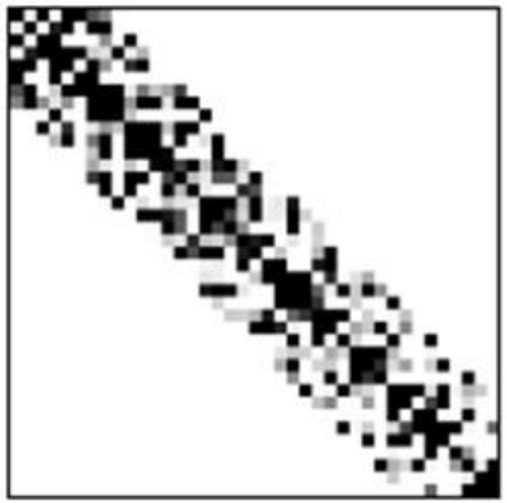
Sparse banded models



Sparse models

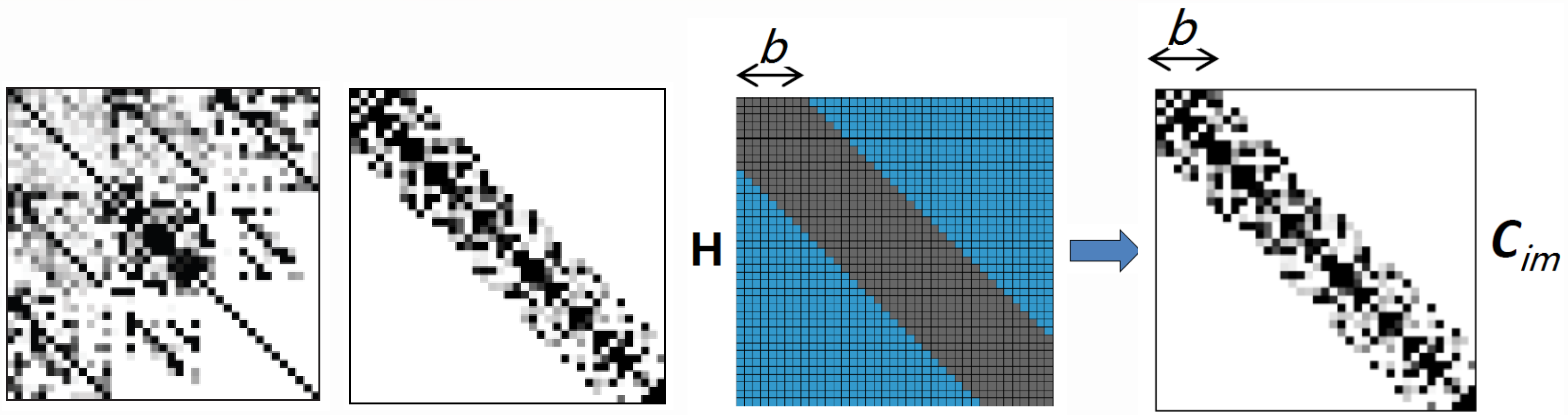


Sparse models
feature reorder

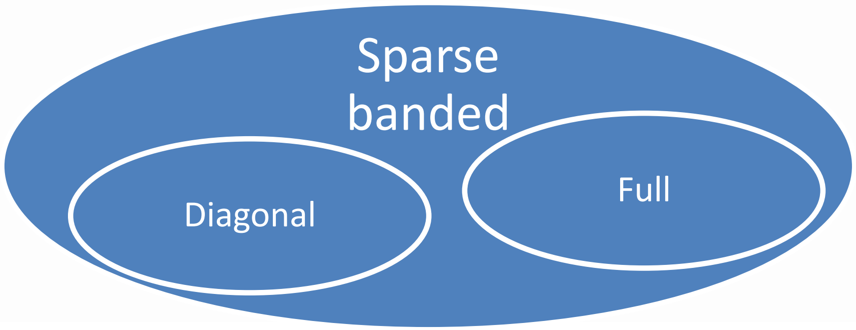


Sparse banded
models

Training of sparse banded models



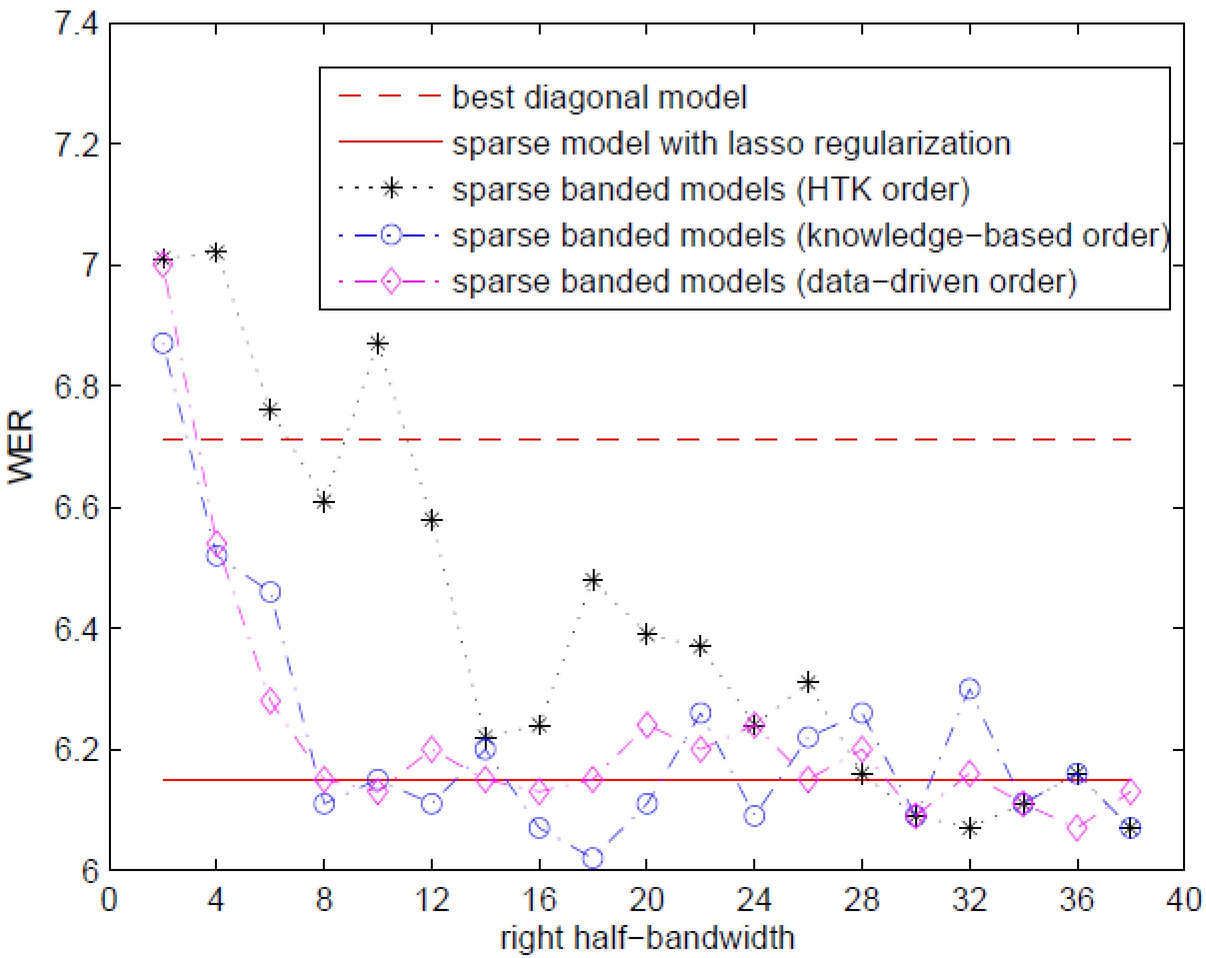
- Weighted lasso: $f(\boldsymbol{C}_{im}) = -\|\boldsymbol{H} * \boldsymbol{C}_{im}\|_1$
- $\boldsymbol{H}(k,l) = \infty \Longrightarrow \boldsymbol{C}_{im}(k,l) = 0$



Importance of the feature order

- $\boldsymbol{O} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}); \boldsymbol{C} = \boldsymbol{\Sigma}^{-1}; \boldsymbol{C}_{ij} = 0 \Longrightarrow o_i$ and o_j are conditionally independent (CI), given other variables.
- Rearrange the feature order so that o_i and o_j are CI if $|i - j| > b$
- Three orders are investigated:
  - HTK order : $m_1 \cdots m_{13} \Delta m_1 \cdots \Delta m_{13} \Delta\Delta m_1 \cdots \Delta\Delta m_{13}$
  - Knowledge-based order : $m_1 \Delta m_1 \Delta\Delta \mathrm{m}_1 \cdots m_{13} \Delta \mathrm{m}_{13} \Delta\Delta m_{13}$
  - Data-driven order : $m_1 \Delta\Delta m_1 \cdots \Delta m_6 \Delta m_{10}$

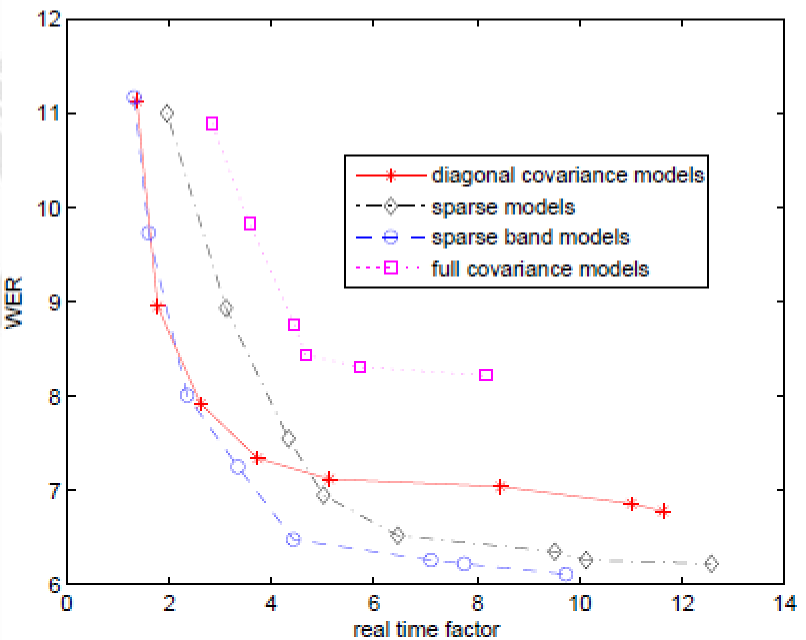
Results on the development data

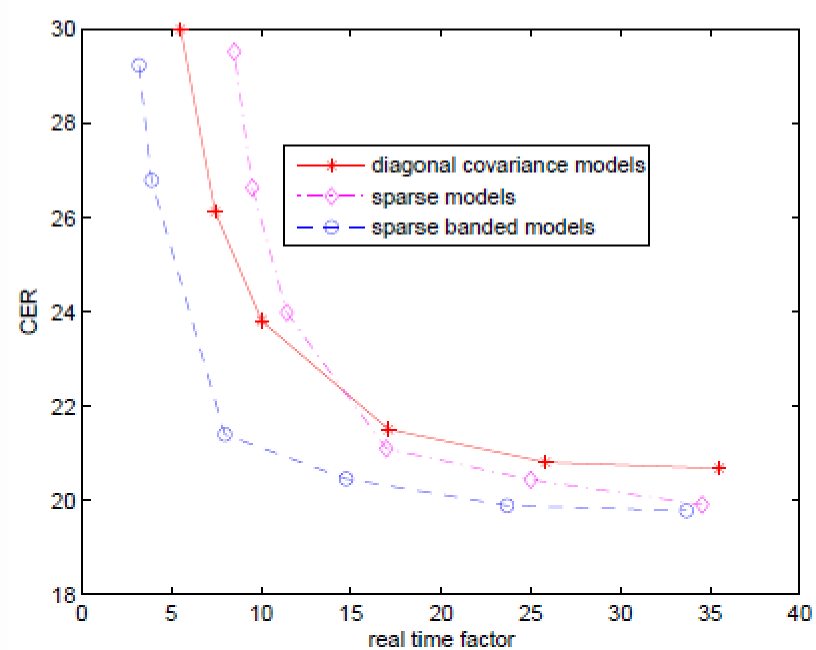


Results on the test data

Model type	#Gaussians	WER	Rel. improv.	Significant?
Full	2	10.5	-7.1	No
Diagonal	10	9.84	----	----
Sparse	4	**8.77**	10.9%	Yes
Band8	4	8.91	9.5	Yes

Decoding time





- Sparse banded modes are the fastest since: 1) smaller searching beam-widths; 2) less model parameters.

Model	#Gaussian components	#total model parameters
diagonal	10	2,491,090
full	1	2,580,719
sparse	2	5,169,440
band8	2	2,041,898

Discriminative training

- MMI objective function:

$$\hat{\Theta} = \underset{\Theta}{Argmax}\{\log P(\boldsymbol{w}_r|\boldsymbol{O},\boldsymbol{\Theta})\}$$

- New Objective function

$$L(\boldsymbol{\Theta}) = \log P(\boldsymbol{w}_r|\boldsymbol{O},\boldsymbol{\Theta}) - \sum_{i=2}^{S-1}\sum_{m=1}^{M}\rho||\boldsymbol{C}_{im}||_1$$

- A valid weak-sense auxiliary function is

$$\begin{aligned}\mathrm{Q}(\boldsymbol{\Theta};\boldsymbol{\Theta}')=&\mathrm{Q}^{\mathrm{n}}(\boldsymbol{\Theta};\boldsymbol{\Theta}')\text{-}\mathrm{Q}^{\mathrm{d}}(\boldsymbol{\Theta};\boldsymbol{\Theta}') && \text{Same as ML training}\\ &+\mathrm{Q}^{\mathrm{s}}(\boldsymbol{\Theta};\boldsymbol{\Theta}') && \text{Ensure stability}\\ &+\mathrm{Q}^{\mathrm{I}}(\boldsymbol{\Theta};\boldsymbol{\Theta}') && \text{Improve generalization}\\ &-\sum_{i=2}^{S-1}\sum_{m=1}^{M}\rho||\boldsymbol{C}_{im}||_1 && \text{Regularization term}\end{aligned}$$

Results on the WSJ testing data

Model type	#Gaussians	ML training	MMI
Full	2	11.68	9.18
Diagonal	10	9.84	9.04
Diagonal+ STC	10	9.26	8.66
Sparse	4	8.55	8.05

Summary

- Sparse models are effective in dealing with the problems that conventional diagonal and full covariance models face: computation, incorrect model assumptions and over-fitting when training data is insufficient.
- We derive the overall training process under the HMM framework using both maximum likelihood training and discriminative training.
- The proposed sparse models subsume the traditional diagonal and full covariance models as special cases.

Thank you!