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
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

- 1950: First ASR (AUDREY)
- 1960: Dynamic Time Warping (DTW)
- 1970: Hidden Markov Models
- 1980: Neural Network
- 2000: Deep Neural Network
• 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

\[ \hat{W} = \arg \max_w P(W | O) = \arg \max_w \frac{P(O | W)P(W)}{P(O)} = \arg \max_w P(O | W)P(W) \]

Diagram with steps:
1. Feature Extraction
2. Decoder
3. "How are you"
4. Acoustic Model
5. Dictionary
6. Language Model
Statistical speech recognition system

- **Language Model:**
  - $P(\text{“recognize speech”}) \gg P(\text{“wreck a nice beach”})$

- **Dictionary:**
  - Wreck: r e k
  - Beach: b i th

- **Acoustic Model:**
  - $P(O|\text{“recognize speech”})$
Statistical speech recognition system

\[ \hat{W} = \arg\max_{w} P(W | O) = \arg\max_{w} \frac{P(O | W)P(W)}{P(O)} = \arg\max_{w} P(O | W)P(W) \]

![Diagram of statistical speech recognition system](image)
Acoustic modeling

- *Left-to-right* hidden Markov models (HMMs)
- *GMM-HMM* based acoustic models

\[
p(o_t|s_j) = \sum_m c_{jm} N(o_t; \mu_{jm}, \Sigma_{jm})
\]

\[
\Theta = \{a_{ij}, b_j(o_t)\} = \{a_{ij}, c_{jm}, u_{jm}, \Sigma_{jm}\}
\]
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
Sparse banded inverse covariance matrices (*sparse models*)

- Alleviate over-fitting
- Less computation
- Reasonable model assumption (decorrelated features)

\[
\frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
\]

<table>
<thead>
<tr>
<th>parameter type</th>
<th>number of parameters</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>transitions</td>
<td>1100</td>
<td>~ 0</td>
</tr>
<tr>
<td>weights</td>
<td>5686</td>
<td>0.2</td>
</tr>
<tr>
<td>means</td>
<td>221,754</td>
<td>4.7</td>
</tr>
<tr>
<td>precision matrices</td>
<td>4,435,080</td>
<td>95.1</td>
</tr>
<tr>
<td>total</td>
<td>4,663,620</td>
<td>100</td>
</tr>
</tbody>
</table>
ML training of sparse models

• Maximum likelihood (ML) training
  \[ \hat{\Theta} = \arg \max_{\Theta} \{ \log(P(O|\Theta)) \} \]

• The proposed new objective function
  \[ L(\Theta) = \log(P(O|\Theta)) - \sum_{i=2}^{S-1} \sum_{m=1}^{M} \rho \| C_{im} \|_1 \]

• Auxiliary function:
  \[ Q(\Theta; \Theta') = \sum_{q} \sum_{m} P(q, m|\Theta', O) \log(P(q, m, O|\Theta)) - \sum_{i=2}^{S-1} \sum_{m=1}^{M} \rho \| C_{im} \|_1 \]

• Properties of the auxiliary function:
  \[ L(\Theta) - L(\Theta') \geq Q(\Theta; \Theta') - Q(\Theta'; \Theta') \]
Maximizing the auxiliary function

\[
\max_\Theta Q(\Theta; \Theta')
\]

\[
P(q, m, o|\Theta) = \prod_{t=1}^{T} a_{qt} a_{qt+1} c_{qm} b_{qm} (o_t)
\]

Forward and backward probabilities
Conditional independent assumptions of HMM

- The precision matrices can be updated using

\[
\tilde{C}_{im} = \arg\max_{C_{im} > 0} \{ \log\det C_{im} - \text{trace}(S_{im} C_{im}) - \lambda ||C_{im}||_1 \}
\]

- \( \lambda = \frac{2\rho}{\gamma_{im}} \) and \( S_{im} \) 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 (39³ triphones)
  – 2843 tied HMM states
Tuning results on the dev. 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.

<table>
<thead>
<tr>
<th>Model type</th>
<th>#Gaussians</th>
<th>WER</th>
<th>Rel. improv.</th>
<th>Significant?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>2</td>
<td>10.5</td>
<td>-7.1</td>
<td>No</td>
</tr>
<tr>
<td>Diagonal</td>
<td>10</td>
<td>9.84</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Sparse</td>
<td>4</td>
<td>8.77</td>
<td>10.9%</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Sparse banded models

Sparse models

Sparse models

Sparse banded models
Training of sparse banded models

- Weighted lasso: $f(C_{im}) = -\|H \ast C_{im}\|_1$
- $H(k, l) = \infty \implies C_{im}(k, l) = 0$
Importance of the feature order

- \( \mathbf{O} \sim N(\mu, \Sigma); \mathbf{C} = \Sigma^{-1}; \mathbf{C}_{ij} = 0 \Rightarrow o_i \) 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: \( m_1 \Delta m_1 \Delta \Delta m_1 \cdots m_{13} \Delta m_{13} \Delta \Delta m_{13} \)
  - Data-driven order: \( m_1 \Delta \Delta m_1 \cdots \Delta m_6 \Delta m_{10} \)
Results on the development data

The diagram shows the WER (Word Error Rate) as a function of the right half-bandwidth. Various models are compared, including:
- best diagonal model
- sparse model with lasso regularization
- sparse banded models (HTK order)
- sparse banded models (knowledge-based order)
- sparse banded models (data-driven order)
### Results on the test data

<table>
<thead>
<tr>
<th>Model type</th>
<th>#Gaussians</th>
<th>WER</th>
<th>Rel. improv.</th>
<th>Significant?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>2</td>
<td>10.5</td>
<td>-7.1</td>
<td>No</td>
</tr>
<tr>
<td>Diagonal</td>
<td>10</td>
<td>9.84</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Sparse</td>
<td>4</td>
<td>8.77</td>
<td>10.9%</td>
<td>Yes</td>
</tr>
<tr>
<td>Band8</td>
<td>4</td>
<td>8.91</td>
<td>9.5</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Decoding time

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

<table>
<thead>
<tr>
<th>Model</th>
<th>#Gaussian components</th>
<th>#total model parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>diagonal</td>
<td>10</td>
<td>2,491,090</td>
</tr>
<tr>
<td>full</td>
<td>1</td>
<td>2,580,719</td>
</tr>
<tr>
<td>sparse</td>
<td>2</td>
<td>5,169,440</td>
</tr>
<tr>
<td>band8</td>
<td>2</td>
<td>2,041,898</td>
</tr>
</tbody>
</table>
Discriminative training

- MMI objective function:
  \[ \hat{\Theta} = \arg\max_\Theta \{ \log P(w_r|O, \Theta) \} \]

- New Objective function
  \[ L(\Theta) = \log P(w_r|O, \Theta) - \sum_{i=2}^{S-1} \sum_{m=1}^M \rho ||C_{im}||_1 \]

- A valid weak-sense auxiliary function is
  \[ Q(\Theta; \Theta') = Q^n(\Theta; \Theta') - Q^d(\Theta; \Theta') + Q^s(\Theta; \Theta') + Q^I(\Theta; \Theta') \]
  \[ - \sum_{i=2}^{S-1} \sum_{m=1}^M \rho ||C_{im}||_1 \]

  - Same as ML training
  - Ensure stability
  - Improve generalization
  - Regularization term
## Results on the WSJ testing data

<table>
<thead>
<tr>
<th>Model type</th>
<th>#Gaussians</th>
<th>ML training</th>
<th>MMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>2</td>
<td>11.68</td>
<td>9.18</td>
</tr>
<tr>
<td>Diagonal</td>
<td>10</td>
<td>9.84</td>
<td>9.04</td>
</tr>
<tr>
<td>Diagonal+ STC</td>
<td>10</td>
<td>9.26</td>
<td>8.66</td>
</tr>
<tr>
<td>Sparse</td>
<td>4</td>
<td>8.55</td>
<td>8.05</td>
</tr>
</tbody>
</table>
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!