Structured SVM for Continuous Speech Recognition

Shi-Xiong Zhang

with Mark Gales

Microsoft, Bellevue, November 25, 2013
Outline

Motivation
   Review Generative and Discriminative Models

Structured SVM for Speech Recognition
   Joint Features
   Training & Decoding
   Related to: HCRF, SCRF, Hybrid system

Experiments

Future Work
PART I: Motivation
Outline

Motivation
Review Generative and Discriminative Models

Structured SVM for Speech Recognition
Joint Features
Training & Decoding
Related to: HCRF, SCRF, Hybrid system

Experiments

Future Work
Speech Recognition: given observation sequence $O$
find “most likely” word sequence $w$

$$\max_w P(w|O)$$

Classification requires class posteriors

- **Generative model** – through Bayes rule

  $$P(w|O) = \frac{p(O|w) P(w)}{p(O)}$$

- **Discriminative model** – directly model posterior

  $$P(w|O; \alpha)$$
Generative Models: HMMs

model likelihood function \(p(O|w)\)

- different length between \(O\) and \(w\) \(\Rightarrow\) introduce latent variable: states

Hidden Markov models:

- Markov assumption: transition probability \(P(s_t|s_{t-1})\)
- emission probability \(p(o_t|s_t)\)
Generative Models: HMMs

model likelihood function $p(O|w)$

- different length between $O$ and $w \Rightarrow$ introduce latent variable: states

Hidden Markov models:

- Markov assumption: transition probability $- P(s_t|s_{t-1})$
  : emission probability $- p(o_t|s_t)$ (GMM/DNN)
Discriminative Models: HCRF

Hidden Conditional Random Fields:

– allow general transition features – \( \phi(s_{t-1}, s_t) \)
– emission features – \( \phi(o_t, s_t) \)

\[
P(w|O; \alpha) = \frac{1}{Z} \sum_s \exp \left( \sum_t \alpha^T \begin{bmatrix} \phi(o_t, s_t) \\ \phi(s_{t-1}, s_t) \end{bmatrix} \right)
\]
Segmental Conditional Random Fields:

- Allow long-term dependency in observations \( \Rightarrow \phi(O_{\tau:t}, w_i) \)

\[
P(w|O; \alpha) = \frac{1}{Z} \sum_s \exp \left( \sum_i \alpha^T \begin{bmatrix} \phi(O_{\tau:t}, w_i) \\ \phi(w_{i-1}, w_i) \end{bmatrix} \right)
\]

directly model posterior with segment-level features:
## Discriminative Training

<table>
<thead>
<tr>
<th>Models</th>
<th>MMI</th>
<th>MPE</th>
<th>Max Margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HCRF/SCRF</td>
<td>✓</td>
<td>✓</td>
<td>☹</td>
</tr>
</tbody>
</table>
### Discriminative Training

<table>
<thead>
<tr>
<th>Models</th>
<th>MMI</th>
<th>MPE</th>
<th>Max Margin</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMMs</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>HCRF/SCRF</td>
<td>✔</td>
<td>✔</td>
<td>☹</td>
</tr>
</tbody>
</table>

**Motivation 1: Max Margin Discriminative Models**
Motivation 2: Extending SVM to CSR

SVMs directly generate boundaries between classes, $\alpha^T \phi(O)$

apply to speech?
Motivation 2: Extending SVM to CSR

SVMs directly generate boundaries between classes, $\alpha^T \phi(O)$

apply to speech?

- map variable length $O$ to fixed dimensional $\phi(O)$
- 1-v-1 voting (binary classification)
Motivation 2: Extending SVM to CSR

SVMs directly generate boundaries between classes, $\alpha^T \phi(O)$

apply to speech?

- For isolated digit recognition, no problem (10 classes)
- For continuous speech? too many classes! (6 digits $\Rightarrow 10^6$ classes)
Motivation 2: Extending SVM to CSR

Simplest approach

step1 Using HMM segmentation

step2 For each segment, apply SVMs
Motivation 2: Extending SVM to CSR

Simplest approach

Step 1: Using HMM segmentation

Step 2: For each segment, apply SVMs

Problem: Treat each segment independently
Motivation 2: Extending SVM to CSR

Simplest approach

\[ \text{FOUR} \quad \text{ONE} \quad \text{SEVEN} \]

- Using HMM segmentation
- For each segment, apply SVMs

Problem: Treat each segment independently

Alternative approach

- Can we extend SVM to model whole continuous speech?
  \[ \Rightarrow \text{Structured SVM!} \]
Extend SVMs to Classify Utterances?

— Infinite number of classes (Is this possible?)
Recall other Sequence Models

<table>
<thead>
<tr>
<th>Models</th>
<th>Isolated ( w ) → Sequence ( w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generative</td>
<td></td>
</tr>
<tr>
<td>↓</td>
<td></td>
</tr>
<tr>
<td>Discriminative</td>
<td></td>
</tr>
<tr>
<td>↓</td>
<td></td>
</tr>
<tr>
<td>Discriminate</td>
<td></td>
</tr>
</tbody>
</table>
## Summary of Sequence Models

<table>
<thead>
<tr>
<th>Training</th>
<th>Isolated $w$</th>
<th>Sequence $w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>Naive Bayes $p(O</td>
<td>w)$</td>
</tr>
</tbody>
</table>

![Diagram of sequence models]
## Summary of Sequence Models

<table>
<thead>
<tr>
<th>Training</th>
<th>Isolated $w$</th>
<th>Sequence $w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>Naive Bayes $p(O</td>
<td>w)$</td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>CML</td>
<td>Logistic $P(w</td>
<td>O)$</td>
</tr>
</tbody>
</table>

Frame features

![](image)

$O_1$, ..., $O_t$, ..., $O_T$
## Summary of Sequence Models

<table>
<thead>
<tr>
<th>Training</th>
<th>Isolated $w$</th>
<th>$\rightarrow$</th>
<th>Sequence $w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>Naive Bayes $p(O</td>
<td>w)$</td>
<td>$\rightarrow$</td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
<td>$\rightarrow$</td>
<td>↓</td>
</tr>
<tr>
<td>CML</td>
<td>Logistic $P(w</td>
<td>O)$</td>
<td>$\rightarrow$</td>
</tr>
</tbody>
</table>

### Segmental features

... $o_{t-1}$ $o_t$ $o_{t+1}$ ...

... $o_1$ $o_t$ $o_T$ ...
Where does the Structured SVM sit?

Map of All Models
Where does the Structured SVM sit?

Map of All Models

<table>
<thead>
<tr>
<th>Training</th>
<th>Isolated $w$</th>
<th>→</th>
<th>Sequence $w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>Naive Bayes $p(O</td>
<td>w)$</td>
<td>→</td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>CML</td>
<td>Logistic $P(w</td>
<td>O)$</td>
<td>→</td>
</tr>
<tr>
<td>↓</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>Max Margin</td>
<td>SVM $\alpha_w^T\phi(O)$</td>
<td>→</td>
<td></td>
</tr>
</tbody>
</table>
Where does the Structured SVM sit?

Map of All Models

<table>
<thead>
<tr>
<th>Training</th>
<th>Isolated $w$</th>
<th>$\rightarrow$</th>
<th>Sequence $w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>Naive Bayes $p(O</td>
<td>w)$</td>
<td>$\rightarrow$</td>
</tr>
<tr>
<td>↓</td>
<td>$\downarrow$</td>
<td></td>
<td>$\downarrow$</td>
</tr>
<tr>
<td>CML</td>
<td>Logistic $P(w</td>
<td>O)$</td>
<td>$\rightarrow$</td>
</tr>
<tr>
<td>↓</td>
<td>$\downarrow$</td>
<td></td>
<td>$\downarrow$</td>
</tr>
<tr>
<td>Max Margin</td>
<td>SVM $\alpha^T_w\phi(O)$</td>
<td>$\rightarrow$</td>
<td>Structured SVM</td>
</tr>
</tbody>
</table>
PART II: Structured SVM
Outline

Motivation
Review Generative and Discriminative Models

Structured SVM for Speech Recognition
Joint Features
Training & Decoding
Related to: HCRF, SCRF, Hybrid system

Experiments

Future Work
Motivation: Extending SVM to Utterance Level

Model whole utterance – observations $O$, word sequence $w$

$\alpha^T_w \phi(O)$
Motivation: Extending SVM to Utterance Level

Model whole utterance – observations $O$, word sequence $w$
Model whole utterance – observations $O$, word sequence $w$

- *joint* features $\phi(O, w)$  How to build?

- Decoding by match score: $\arg\max_w \alpha^T \phi(O, w)$
Structured SVM for Speech Recognition

Joint Features

Model whole utterance – observations $O$, word sequence $w$

- joint features $\phi(O, w)$
Structured SVM for Speech Recognition

*Joint Features*

Joint features $\phi(O, w)$

How to extract local features?
### Structured SVM for Speech Recognition

#### Joint Features

**Joint features** \( \phi(O, w) \)  

**Possible local features**

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame-level</td>
<td>Gaussian sufficient statistics [2]</td>
</tr>
<tr>
<td></td>
<td>MLP posteriors [10]</td>
</tr>
<tr>
<td></td>
<td>bottleneck features [1]</td>
</tr>
<tr>
<td>Segment-level</td>
<td>log likelihood features [9]</td>
</tr>
<tr>
<td></td>
<td>derivative features [6]</td>
</tr>
<tr>
<td></td>
<td>event detectors [11]</td>
</tr>
</tbody>
</table>
Structured SVM for Speech Recognition

Joint Features

joint features $\phi(O, w)$

generative models based features

$$
\log p(O; \lambda^{\text{one}}) \\
\vdots \\
\log p(O; \lambda^{\text{zero}})
$$
Structured SVM for Speech Recognition

Joint Features

Joint features $\phi(O, w)$

Generative models based features

$\begin{pmatrix}
\log p(O; \lambda^{\text{one}}) \\
\vdots \\
\log p(O; \lambda^{\text{zero}})
\end{pmatrix}$

- Smiley face: generative + discriminative models
- Smiley face: model compensation
- Smiley face: joint learn with language model features
Structured SVM for Speech Recognition

Joint Features

Joint features $\phi(O, w)$

Decoding by match score: $\arg \max_w \alpha^T \phi(O, w)$
Training
Structured SVM for Speech Recognition

Training

Training $\alpha$, Maximize the Margin
Subject to: score of correct $w_{ref}$ $\geq$ all competing $w$
Structured SVM for Speech Recognition

Training $\alpha$, Maximize the Margin

Subject to: score of correct $w_{\text{ref}} \geq$ all competing $w$

$$\min \frac{1}{2} ||\alpha||^2$$

S.T.

$$\alpha^T \phi(\ldots, \text{"Hi there"}) \geq \alpha^T \phi(\ldots, \text{"Hi Dog"}) + 1$$

$$\vdots$$

$$\min \frac{1}{2} ||\alpha||^2$$

S.T.

$$\alpha^T \phi(\ldots, \text{"See you"}) \geq \alpha^T \phi(\ldots, \text{"Saw you"}) + 1$$

$$\vdots$$
To generalize the training

- Replace “0-1 loss” as $\mathcal{L}(w_{\text{ref}}, w)$
- Introduce slack variable $\xi_i$

$$\min \frac{1}{2} \| \alpha \|^2 + C \sum_{i=1}^{n} \xi_i$$

S.T. for each training utterance $i$

$$\alpha^T \phi(\text{"Hi there"}) \geq \max_{\text{words}} \left\{ \alpha^T \phi(\text{words}) + \text{loss} \right\} - \xi_i$$
Structured SVM for Speech Recognition

Training

\[
\min \frac{1}{2} \|\alpha\|^2 + C \sum_{i=1}^{n} \xi_i
\]

S.T. for each training utterance \(i\)

\[
\alpha^T \phi(O^{(i)}, w_{\text{ref}}) \geq \max_{\text{words}} \left\{ \alpha^T \phi(O^{(i)}, w) + \text{loss} \right\} - \xi_i
\]

Unconstrained form

\[
\frac{1}{2} \|\alpha\|^2 + C \sum_{i=1}^{n} \left[ -\alpha^T \phi(O^{(i)}, w_{\text{ref}}) + \max_{w \neq w_{\text{ref}}} \left\{ \alpha^T \phi(O^{(i)}, w) + \text{loss} \right\} \right]
\]
Features depend on the segmentation $s$

For efficiency, consider one “best” segmentation – define “best”?

- Previously, $\phi(O, w) = \phi(O, w; s_{hmm})$
  Problem: Best segmentation in HMM may not the best in SSVM

- What we want: $\max_s \alpha^T \phi(O, w; s)$
Can we optimize segmentation $s$ in Structured SVMs?

- How to learn $\alpha$ and $s$ jointly in training?
- How to find $w$ and $s$ in decoding?
Training with Optimal Segmentation

• Previously, no variable $s$, $\min_\alpha$ (Convex Optimization)

$$\frac{1}{2}\|\alpha\|_2^2 + C \sum_{i=1}^n \left[ -\alpha^T \phi(O^{(i)}, w^{(i)}_{\text{ref}}) + \max_{w \neq w_{\text{ref}}} \left\{ \mathcal{L}(w, w^{(i)}_{\text{ref}}) + \alpha^T \phi(O^{(i)}, w) \right\} \right] +$$

• Now, Optimizing $s$ with $\alpha$ (Nonconvex Optimization)

$$\left[ -\max_s \alpha^T \phi(O^{(i)}, w^{(i)}_{\text{ref}}, s) + \max_{w \neq w_{\text{ref}}, s} \left\{ \mathcal{L}(w, w^{(i)}_{\text{ref}}) + \alpha^T \phi(O^{(i)}, w, s) \right\} \right] +$$
Training with Optimal Segmentation

- Previously, no variable \( s \), \( \min_\alpha \) (Convex Optimization)

\[
\frac{1}{2} ||\alpha||_2^2 + C \sum_{i=1}^{n} \left[ -\alpha^T \phi(O^{(i)}, w^{(i)}_{ref}) + \max_{w \neq w_{ref}} \left\{ \mathcal{L}(w, w^{(i)}_{ref}) + \alpha^T \phi(O^{(i)}, w) \right\} \right]
\]

- Now, Optimizing \( s \) with \( \alpha \) (Nonconvex Optimization)

\[
-\max_{s} \alpha^T \phi(O^{(i)}, w^{(i)}_{ref}, s) + \max_{w \neq w_{ref}, s} \left\{ \mathcal{L}(w, w^{(i)}_{ref}) + \alpha^T \phi(O^{(i)}, w, s) \right\}
\]
Algorithm (EM-style, guaranteed to converge):

1. *given* $\alpha \Rightarrow$ find best segmentation $s$

2. *given* $s \Rightarrow$ original SSVM convex optimisation $\alpha$
Decoding
Decoding — frame-level features

Decoding by match score: \[
\arg \max_{w, s} \alpha^T \phi(O, w, s)
\]

- frame-level feature: \[
\alpha^{ac} \phi(o_t, s_t) + \alpha^{lm} \phi(s_{t-1}, s_t)
\]

\[\phi(o_t, s_t)\]

\[\alpha^T \phi(o_t, s_t)\]

\[\alpha^T \phi(o_t, s_t)\]

\[\alpha^T \phi(o_t, s_t)\]

\[\text{nonlinear map}\]

\[\text{states } s\]

1

2

3

\[t \quad t + 1\]
Decoding — segment-level features

- Decoding by match score: \( \arg \max_{w,s} \alpha^T \phi(O, w, s) \)

- Search based on lattices
Related to Prior Work
Structured SVM ≡ Max Margin Log Linear Model

Example of log-linear models

\[ P(w|O; \alpha) = \frac{1}{Z} \exp(\alpha^T \phi(O, w)) \]

- **Same Decoding:** \( \arg \max_w P(w|O; \alpha) = \arg \max_w \alpha^T \phi(O, w) \)
Structured SVM $\equiv$ Max Margin Log Linear Model

Example of log-linear models

$$P(w|O; \alpha) = \frac{1}{Z} \exp(\alpha^T \phi(O, w))$$

- Becomes Structured SVM in Max Margin Training

Margin: minimum distance between correct $w_{ref}$ and competing $w$

$$\min_{w \neq w_{ref}} \left\{ \log \left( \frac{P(w_{ref}|O; \alpha)}{P(w|O; \alpha)} \right) \right\}$$
Structured SVM $\equiv$ Max Margin Log Linear Model

Example of log-linear models

$$P(w|O; \alpha) = \frac{1}{Z} \exp(\alpha^T \phi(O, w))$$

- Becomes Structured SVM in Max Margin Training

$\alpha$ \text{ Margin: minimum distance between correct } w_{\text{ref}} \text{ and competing } w$

$$\min_{w \neq w_{\text{ref}}} \left\{ \log \left( \frac{P(w_{\text{ref}}|O; \alpha)}{P(w|O; \alpha)} \right) \right\}$$

$$\frac{1}{2} ||\alpha||^2_2 + C \sum_{i=1}^{n} \left[ -\alpha^T \phi(O^{(i)}, w^{(i)}_{\text{ref}}) + \max_{w \neq w_{\text{ref}}} \left\{ \alpha^T \phi(O^{(i)}, w) + \mathcal{L}(w, w^{(i)}_{\text{ref}}) \right\} \right]$$
Structured SVM \equiv \text{Max Margin HCRF/SCRF}

Example of HCRF/SCRF

\[ P(w|O; \alpha) = \frac{1}{Z} \sum_s \exp (\alpha^T \phi(O, w, s)) \]

- Same Decoding: \[ \arg \max_{w,s} P(w|O; \alpha) = \arg \max_{w,s} \alpha^T \phi(O, w, s) \]
Structured SVM ≡ Max Margin HCRF/SCRF

Example of HCRF/SCRF

\[ P(w|O; \alpha) = \frac{1}{Z} \sum_s \exp(\alpha^T \phi(O, w, s)) \]

→ Becomes Structured SVM in Max Margin Training

• Margin: minimum distance between correct \( w_{ref} \) and competing \( w \)

\[
\min_{w \neq w_{ref}} \left\{ \log \left( \frac{P(w_{ref}|O; \alpha)}{P(w|O; \alpha)} \right) \right\}
\]

\[
- \max_s \alpha^T \phi(O^{(i)}, w_{ref}^{(i)}, s) + \max_{w \neq w_{ref}, s} \left\{ \alpha^T \phi(O^{(i)}, w, s) + \mathcal{L}(w, w_{ref}^{(i)}) \right\}
\]
Related to Hybrid System

Structured SVM

\[ \phi(o_t, s_t) \]

\[ \alpha^T \phi(o_t, s_t) \]

\[ \alpha^T \phi(o_t, s_t) \]

\[ \alpha^T \phi(o_t, s_t) \]

Hybrid System

\[ P(s_t = 3|o_t) \]

\[ P(s_t = 2|o_t) \]

\[ P(s_t = 1|o_t) \]
Practical Issues
Optimize segmentation

- efficient feature extraction
Practical Issues

Optimize segmentation

✔ efficient feature extraction

Joint feature space very large

✔ Parameter tying – for discriminative models
Practical Issues

Optimize segmentation

- efficient feature extraction

Joint feature space very large

- Parameter tying – for discriminative models

![Diagram of discriminative acoustic model tree and generative model tree](image-url)
Practical Issues

Optimize segmentation

✓ efficient feature extraction

Joint feature space very large

✓ Parameter tying – for discriminative models
✓ Kernelization – avoid explicitly computing feature space

\[
P(w|O; \alpha) = \frac{1}{Z} \exp\left( \alpha^T \phi(O, w) \right)
\]

\[\Rightarrow\]

\[
P(w|O; \beta) = \frac{1}{Z} \exp\left( \sum_i \beta_i k ((O_i, w_i), (O, w)) \right)
\]
Optimize segmentation

✓ efficient feature extraction

Joint feature space very large

✓ Parameter tying – for discriminative models
✓ Kernelization

Efficiency

✓ $n$-slack $\Rightarrow$ 1-slack ($\xi = \sum_{i=1}^{n} \xi_i$)
Optimize segmentation

- efficient feature extraction

Joint feature space very large

- Parameter tying – for discriminative models
- Kernelization

Efficiency

- $n$-slack $\Rightarrow$ 1-slack
- Modifying Prior – $P(\alpha)$

\[
\log P(\alpha) = \log (\mathcal{N}(\alpha; 0, I)) \propto \frac{1}{2} ||\alpha||^2
\]
Practical Issues

Optimize segmentation

✓ efficient feature extraction

Joint feature space very large

✓ Parameter tying – for discriminative models
✓ Kernelization

Efficiency

✓ $n$-slack $\Rightarrow$ 1-slack
✓ Modifying Prior – $P(\alpha)$
✓ Parallelization
Practical Issues

Optimize segmentation

✓ efficient feature extraction

Joint feature space very large

✓ Parameter tying – for discriminative models
✓ Kernelization

Efficiency

✓ $n$-slack $\Rightarrow$ 1-slack
✓ Modifying Prior – $P(\alpha)$
✓ Parallelization
✓ Caching and Pruning
Implementation

Training Phase
Decoding Phase

Generating Numerator Lattice → Search reference segmentation
Generating Denominator Lattice → Search competing \( \mathbf{W} \) and segmentation

Generating Numerator Lattice → Search best \( \mathbf{W} \) and segmentation

Large Margin Training \( \mathbf{\alpha} \)

\( \mathbf{W}^{(i)} \) ref → 1
\( \mathbf{O}^{(i)} \) → 2

\( \mathbf{\alpha} \) Updated

\( \mathbf{\hat{W}} \)

\( \mathbf{\hat{\alpha}} \)

\( \mathbf{\hat{W}} \) Parallelized
Part III: Experiments
Motivation
Review Generative and Discriminative Models

Structured SVM for Speech Recognition
Joint Features
Training & Decoding
Related to: HCRF, SCRF, Hybrid system

Experiments

Future Work
## Aurora 2: Noise corrupted continuous digit task.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Criteria</th>
<th>Err. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM+VTS</td>
<td>46,732</td>
<td>ML</td>
<td>9.5</td>
</tr>
<tr>
<td>SVM</td>
<td>+144</td>
<td>MM</td>
<td>8.3</td>
</tr>
<tr>
<td>SCRF</td>
<td>+144</td>
<td>CML</td>
<td>8.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MWE</td>
<td>7.9</td>
</tr>
<tr>
<td>Struct SVM</td>
<td>+144</td>
<td>MM</td>
<td>7.4</td>
</tr>
</tbody>
</table>

• Gain over HMMs: 22% in Aurora 2
Experiments

Aurora 2: Noise corrupted continuous digit task.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Criteria</th>
<th>Err. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM+VTS</td>
<td>46,732</td>
<td>ML</td>
<td>9.5</td>
</tr>
<tr>
<td>SVM</td>
<td>+144</td>
<td>MM</td>
<td>8.3</td>
</tr>
<tr>
<td>SCRF</td>
<td>+144</td>
<td>CML</td>
<td>8.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MWE</td>
<td>7.9</td>
</tr>
<tr>
<td>Struct SVM</td>
<td>+144</td>
<td>MM</td>
<td>7.4</td>
</tr>
</tbody>
</table>

Training Algorithm for SSVM

<table>
<thead>
<tr>
<th>Training Algorithm for SSVM</th>
<th>Err. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM segment</td>
<td>7.6</td>
</tr>
<tr>
<td>+ optimal segmentation</td>
<td>7.4</td>
</tr>
<tr>
<td>+ kernel (2nd-order poly)</td>
<td>7.3</td>
</tr>
</tbody>
</table>

• Gain over HMMs: **22% in Aurora 2**
Aurora 4: Noise corrupted medium to large vocabulary task

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Criteria</th>
<th>Err. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM+VTS</td>
<td>3.98M</td>
<td>ML</td>
<td>17.8</td>
</tr>
<tr>
<td>SCRF</td>
<td>+2210</td>
<td>MPE</td>
<td>17.4</td>
</tr>
<tr>
<td>Struct SVM</td>
<td>+2210</td>
<td>MM</td>
<td>16.8</td>
</tr>
</tbody>
</table>
Aurora 4: Noise corrupted medium to large vocabulary task

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Criteria</th>
<th>Err. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM+VTS</td>
<td>3.98M</td>
<td>ML</td>
<td>17.8</td>
</tr>
<tr>
<td>SCRF</td>
<td>+2210</td>
<td>MPE</td>
<td>17.4</td>
</tr>
<tr>
<td>Struct SVM</td>
<td>+2210</td>
<td>MM</td>
<td>16.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Criteria</th>
<th>Err. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM+VTS+SAT</td>
<td>3.98M</td>
<td>ML</td>
<td>16.0</td>
</tr>
<tr>
<td>SCRF</td>
<td>+2210</td>
<td>MPE</td>
<td>15.7</td>
</tr>
<tr>
<td>Struct SVM</td>
<td>+2210</td>
<td>MM</td>
<td>15.4</td>
</tr>
</tbody>
</table>

- Gain over HMMs: 6% in Aurora 4
Aurora 4: Noise corrupted medium to large vocabulary task

- features based on MPE trained HMM

\[ \phi(O) = \begin{bmatrix}
\log p(O|\lambda^{HMM_i}) \\
\vdots
\end{bmatrix} \]

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Criteria</th>
<th>Err. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM+MPE+VTS+SAT</td>
<td>3.98M</td>
<td>MPE</td>
<td>15.3</td>
</tr>
<tr>
<td>SCRF</td>
<td>+2210</td>
<td>MPE</td>
<td>15.3</td>
</tr>
<tr>
<td>Struct SVM</td>
<td>+2210</td>
<td>MM</td>
<td>15.0</td>
</tr>
</tbody>
</table>
Part IV: Future Work
Outline

**Motivation**
Review Generative and Discriminative Models

**Structured SVM for Speech Recognition**
Joint Features
Training & Decoding
Related to: HCRF, SCRF, Hybrid system

**Experiments**

**Future Work**
DNN-SSVM \ VS. DNN-HMM?

DNN frame/segmental features \Rightarrow Hybrid Discriminative Models

\[ \phi(O) = \begin{bmatrix} \log p(O | \lambda_{DNN-HMM_1}) \\ \vdots \\ P_{MLP}(s_t | o_t) \end{bmatrix} \]

\[ \phi(o_t) = \begin{bmatrix} \vdots \\ P_{MLP}(s_t | o_t) \end{bmatrix} \]
Deep Structured SVM?

\[ \phi(o_t, s_t) \]

nonlinear map

\[ \alpha^T \phi(o_t, s_t) \]

\[ \alpha^T \phi(o_t, s_{t-1}) \]

states \( s \)

\[ \alpha^T \phi(s_{t-1}, s_t) \]

\[ o_t \]

\[ t \]

\[ t + 1 \]
Deep Structured SVM?

\[ \alpha^T \phi(o_t, s_t) \]

\[ \alpha^T \phi(s_{t-1}, s_t) \]

\[ \phi_2(\phi_1(\cdot)) \]

nonlinear map

states s

3

2

1

{1, 2, 3}
Deep Structured SVM?

\[ \phi_3 (\phi_2 (\phi_1 (\cdot))) \]

\[ \alpha^T \phi (o_t, s_t) \]

\[ \alpha^T \phi (o_t, s_t) \]

\[ \alpha^T \phi (o_t, s_t) \]

\[ \alpha^T \phi (s_{t-1}, s_t) \]
Structured SVMs ⇔ Max Margin Discriminative Models

Logistic ⟷ HCRF/SCRF

SVM ⟷ Struct SVM

Structured SVM for ASR

- Joint Features
- Training and Decoding
- Practical Issues

Deep Structured SVM?
THANKS!

Q & A
Complementary
Possible Features (Summary)

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Example Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>frame-level</td>
<td></td>
</tr>
<tr>
<td>Gaussian statistics [2]</td>
<td>$^{o_t}_{\text{diag}(o_t o_T)}$</td>
</tr>
<tr>
<td>MLP posterior</td>
<td>$P_{\text{MLP}}(s</td>
</tr>
<tr>
<td>state transition</td>
<td>$\delta(s_t = s) \delta(s_{t+1} = s')$, $\forall s, s'$</td>
</tr>
<tr>
<td>Segmental</td>
<td></td>
</tr>
<tr>
<td>Log Likelihoods [9]</td>
<td>$\log p_{\lambda}(O</td>
</tr>
<tr>
<td>Derivative [6]</td>
<td>$\nabla_{\lambda} \log p_{\lambda}(O</td>
</tr>
<tr>
<td>Event detector [11]</td>
<td>$\psi_{\text{scr}}(O, w_i)$</td>
</tr>
<tr>
<td>word n-gram</td>
<td>$\delta(w_i = w) \delta(w_{i+1} = w')$, $\forall w, w'$</td>
</tr>
</tbody>
</table>
Generative Models: HMMs

GMM-HMM

DNN-HMM [4]
Incorporate structures into SVM classes ⇒ Structured SVM!

What is the structures?

- Sequence structure in Handwriting Recognition

Input \[\begin{array}{c}
| b | r | a | c |
\end{array}\] \rightarrow Output \[\begin{array}{c}
| b | r | a | c |
\end{array}\]
Training $\alpha$ in QP form

$$\min_{\alpha, \xi} \frac{1}{2} \|\alpha\|^2 + \frac{C}{n} \sum_{i=1}^{n} \xi_i$$

s.t. for every utterance $i$, for ALL the competing labels $w^{(i)}$

$$\alpha^T \phi(O^{(i)}, w^{(i)}_{\text{ref}}) - \alpha^T \phi(O^{(i)}, w^{(i)}_{\ast}) \geq \mathcal{L}(w^{(i)}_{\text{ref}}, w^{(i)}_{\ast}) - \xi_i$$

**Handling infinite number of constraints - Cutting Plane Algorithm**
Cutting Plane Algorithm

repeat
    for $i = 1, \ldots, n$ do
        ① constraint set $\leftarrow$ Generate a new constraint ($w^{(i)}_*$)
        ② $\alpha \leftarrow$ Solving QP with current constraint set
    end for
until no new constraints
Decoding with optimal segmentation

- Search optimal position of “black nodes”, and the label between them.

![Diagram showing HMM models with labels and particle weights]

- Viterbi-style search, \( \psi(t) = \max_{t_{st}, w} \left\{ \psi(t_{st}) + \sum_{k=\text{“one”}}^{\text{“zero”}} \alpha_{k}^{(w)} \text{HMM}_k \right\} \)
  - Related to Factorial HMM inference
F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky.
Probabilistic and bottle-neck features for lvcsr of meetings.

A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt.
Hidden conditional random fields for phone classification.
In International Conference on Speech Communication and Technology, 2005.

G. Heigold, R. Schluter, and H. Ney.
On the equivalence of gaussian hmm and gaussian hmm-like hidden conditional random fields.


Cutting-plane training of structural svms.

A. Ragni and M. J. F. Gales.
Derivative kernels for noise robust ASR.
S.-X. Zhang and M. J. F. Gales.
Kernelized log linear models for continuous speech recognition.

S.-X. Zhang and M. J. F. Gales.
Structured SVMs for automatic speech recognition.

S.-X. Zhang, A. Ragni, and M. J. F. Gales.
Structured log linear models for noise robust speech recognition.

On using mlp features in lvcsr.

G. Zweig and P. Nguyen.
A segmental CRF approach to large vocabulary continuous speech recognition.
In *ASRU*, 2009.