Discriminative Training for Automatic Speech Recognition

Robert Viehauser

22nd April 2013

Advanced Signal Processing Seminar
Discriminative Training for Automatic Speech Recognition
Heigold, G.; Ney, H.; Schluter, R.; Wiesler, S.

Covered Topics

- Statistical Speech Recognition
- Discriminative Training Criteria
- Parameter Models
- Optimisation
- Implementation
- Experimental Results
- Summary and Outlook
4 Components of an LVCSR System

from [6]
Features

### Usable with discriminative Training
- Short-term power spectrum
- Mel frequency cepstral coefficients (MFCC)
- Perceptual linear prediction (PLP)
- Further enhancement methods (LDA,...)

### Other Approaches
- Feature Extraction through Neural Networks
Problem

**Given**

Sequence of Feature-Vectors $x_1^T$

**Find**

Sequence of Words/Phonemes $w_1^N$

**Statistical Model**

$$\left[ w_1^T \right]_{opt} = \arg\max_{w_1^N} p(w_1^N | x_1^T)$$

Robert Viehauser

Discriminative Training for Automatic Speech Recognition
Problem

Statistical Model cont.

\[
\begin{align*}
\begin{bmatrix} w_1^T \end{bmatrix}_{\text{opt}} &= \arg\max_{w_1^N} p(w_1^N | x_1^T) \\
&= \arg\max_{w_1^N} \frac{p(x_1^T, w_1^N)}{p(x_1^T)} \\
&\text{indep. of } w_1^N \\
&= \arg\max_{w_1^N} p(x_1^T, w_1^N) \\
&= \arg\max_{w_1^N} p(x_1^T | w_1^N) \cdot p(w_1^N)
\end{align*}
\]
Statistical Model cont.

\[
\begin{bmatrix} w_1^T \end{bmatrix}_{opt} = \arg\max_{w_1^N} p(x_1^T | w_1^N) \cdot \underbrace{p(w_1^N)}_{\text{Acoustic Model}} \cdot \underbrace{p(w_1^N)}_{\text{Language Model}}
\]
Standard Acoustic Model

Gaussian Hidden Markov Model

\[
p(x_1^T | w_1^N) = \sum_{S_1^T \in HMM(w_1^N)} p(x_1^T, s_1^T)
\]

\[
= \sum_{S_1^T \in HMM(w_1^N)} \prod_{t=1}^{T} p(x_t | s_t) \cdot p(s_t | s_{t-1})
\]

\[
= \sum_{S_1^T \in HMM(w_1^N)} \prod_{t=1}^{T} p(s_t | s_{t-1}) \sum_{l=1}^{L_s} c_{s_t, l} \mathcal{N}(x_t | \mu_{s_t, l}, \Sigma_{s_t, l})
\]

Parameterset for GHMM: \( p(s_1^T | s_0^{T-1}), c_{s_1^T, l_1^{L_s}}, \mu_{s_1^T, l_1^{L_s}}, \Sigma_{s_1^T, l_1^{L_s}} \rightarrow \Lambda \)
Maximum Likelihood

Standard Approach

Find the most probable parameters for the model given the training data set \((x_1^T, w_1^N)\).

Training Criterion

\[ F(\Lambda) = \log p_{\Lambda}(x_1^T, w_1^N) \]
Maximum Mutual Information (MMI)

Motivation
- Maximizes directly the posterior probability
- Takes also all competing sentences $\tilde{w}_1^N$ into account

Training Criterion

$$F(\Lambda) = \log p_\Lambda(w_1^N|x_1^T) = \log \frac{p_\Lambda(x_1^T, w_1^N)}{\sum_{\tilde{w}_1^N} p_\Lambda(x_1^T, \tilde{w}_1^N)}$$
Maximum Mutual Information (MMI)

MMI vs. ML

from [1]
Minimum Classification Error (MCE)

Motivation
- Assumption that GMMs are not the real distribution
- Aims to maximize the classification error (WER)
- Derived by minimizing the expected loss

Training Criterion

\[ F(\Lambda) = \sigma_\beta \left( \log \frac{p_\Lambda(x^T_1, w^N_1)}{\sum_{\tilde{w}^N_1 \neq w^N_1} p_\Lambda(x^T_1, \tilde{w}^N_1)} \right) \]
Minimum Phone Error (MPE)

Motivation
- Similar motivation as for MCE
- Aims to minimize the phone error rate (Levenshtein-Distance)
- Hypotheses weighted by phone accuracy $A(\tilde{w}_1^N, w_1^N)$

Training Criterion

$$F(\Lambda) = \sum_{\tilde{w}_1^N} \sum_{\tilde{w}_1^N} p_\Lambda(\tilde{w}_1^N | x_1^T) A(\tilde{w}_1^N, w_1^N)$$
l-smoothing

Overfitting Problem

- Discriminative training criteria prone to overfitting
- Critical on less training data
- Introducing a prior to each Gaussian based on the ML statistics
- Essentially for MCE/MPE
Focus on decision boundary

- Generalization reached through closest training samples to the decision boundary (Margin, SVM)
- Focus more on these samples by adding a term to the training criterion: $\exp(-A(\tilde{w}_1^N, w_1^N))$
- Applied on MMI: equal to boosted MMI (BMMI)
Language Model Scale

HMM

\[ p(x_1^T | w_1^N) = \sum_{S_1^T \in HMM(w_1^N)} p(x_1^T, s_1^T)^{1/\kappa} \]

Language Model

\[ p(w_1^N) = p(w_1^N)^\kappa \]

\( \kappa \): language model scale
Feature Transform

Features containing more information

Each training criterion can also be used in the feature space. For instance:

\[ y_t = x_t + M \cdot h_t \]
Comparison

Unified Training Criterion

\[ F(\Lambda) = f \left( \frac{\sum \tilde{w}_1^N p_{\Lambda}(x_1^T, \tilde{w}_1^N) A(\tilde{w}_1^N, w_1^N)}{\sum \tilde{w}_1^N p_{\Lambda}(x_1^T, \tilde{w}_1^N) B(\tilde{w}_1^N, w_1^N)} \right) \]

MMI setup

\[
\begin{align*}
  f(x) &= \log(x) \\
  A(\tilde{w}_1^N, w_1^N) &= \delta(\tilde{w}_1^N, w_1^N) \\
  B(\tilde{w}_1^N, w_1^N) &= 1
\end{align*}
\]
Comparison

Binary Classification

from [1]
Extended Baum-Welch

Strong-sense auxiliary function

\[ G(\lambda, \lambda') - G(\lambda', \lambda') \leq F(\lambda) - F(\lambda') \]

Principe of Expectation Maximization Algorithm

Weak-sense auxiliary function

\[ \frac{\partial}{\partial \lambda} G(\lambda, \lambda')|_{\lambda=\lambda'} = \frac{\partial}{\partial \lambda} F(\lambda)|_{\lambda=\lambda'} \]

Principe of Extended Baum-Welch Algorithm
Rprop

Properties

- Only sign of the partial derivatives is needed
- Separate step size for each parameter
- Simple heuristic
- Good alternative to EBW
- Roughly the same number of iterations as EBW for convergence on conservative initial step size
### Experimental Results

#### EBW vs. Rprop

<table>
<thead>
<tr>
<th>TASK</th>
<th>CRITERION</th>
<th>OPTIMIZATION</th>
<th>TEST WER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIETILL</td>
<td>ML</td>
<td>EM</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>M-MMI</td>
<td>EBW</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RPROP</td>
<td>1.6</td>
</tr>
<tr>
<td>EPPS EN</td>
<td>ML</td>
<td>EM</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td>MPE</td>
<td>EBW</td>
<td>11.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RPROP</td>
<td>11.5</td>
</tr>
<tr>
<td>BNBC CN</td>
<td>ML</td>
<td>EM</td>
<td>17.9</td>
</tr>
<tr>
<td></td>
<td>MPE</td>
<td>EBW</td>
<td>17.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RPROP</td>
<td>16.5</td>
</tr>
</tbody>
</table>

*from [1]*
### Experimental Results

#### Training Criteria

<table>
<thead>
<tr>
<th>EPPS EN</th>
<th>ML</th>
<th>MMI</th>
<th>MCE</th>
<th>MPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEVELOPMENT WER [%]</td>
<td>14.4</td>
<td>13.8</td>
<td>13.8</td>
<td>13.4</td>
</tr>
<tr>
<td>TEST WER [%]</td>
<td>12.0</td>
<td>12.0</td>
<td>11.9</td>
<td>11.5</td>
</tr>
</tbody>
</table>

from [1]
### Experimental Results

#### Margin Term

<table>
<thead>
<tr>
<th>TASK</th>
<th>CRITERION</th>
<th>MARGIN</th>
<th>TEST WER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIETILL</td>
<td>ML</td>
<td>N/A</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>MMI</td>
<td>NO</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YES</td>
<td>1.6</td>
</tr>
<tr>
<td>EPPS EN</td>
<td>ML</td>
<td>N/A</td>
<td>12.0</td>
</tr>
<tr>
<td></td>
<td>MPE</td>
<td>NO</td>
<td>11.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YES</td>
<td>11.3</td>
</tr>
<tr>
<td>BNBC CN</td>
<td>ML</td>
<td>N/A</td>
<td>17.9</td>
</tr>
<tr>
<td></td>
<td>MPE</td>
<td>NO</td>
<td>16.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>YES</td>
<td>16.3</td>
</tr>
</tbody>
</table>

From [1]
References


