An Integrated Framework for Margin-based Sequential Discriminative Training over Lattices using differenced Maximum Mutual Information (dMMI)

Erik McDermott - Google Inc.

September 14, 2012
Error-weighted training using *explicit models of error* (MPE/MWE/sMBR etc.)

*Shifting* of loss function: “margin” (MCE, MPE, bMMI)
- Make shift proportional to error.

Extension of “point” use of margin to integral over margin interval $\rightarrow$ proposal of “*differenced MMI*” ($d$MMI)

$d$MMI: margin & error-dependent loss smoothing/integration
- Unifies margin-modified MMI and MPE
- More general than MPE yet allows a simpler implementation using difference of standard Forward-Backward statistics

Bayesian view & further generalization.
Integrated system optimization

**Composition**

1. **Reference WFSTs**
2. **Reference transcriptions**

**Speech Data**

WFST decoder

- Full competitor WFST or lattice-derived subset
- Cost function
  - Correct
  - Incorrect

**Feedback**

- Language Model
- Dictionary
- Acoustic Model

**Correct**

**Incorrect**

Large-scale discriminative training
Minimum Phone Error
Margin-based variants; bMMI
Non-uniform error for discriminative training

**Word error count**
Reference: 「今日の北海道の天気は晴れです」
Hypothesis: 「昨日の東海道の天気は晴れです」 (Error: 2)

**Phone error count**
Reference: 「kyou no hoko kaidou no ten Nkiwaharedesu」
Hypothesis: 「kino no toku kaidou no ten Nkiwaharedesu」 (Error: 5)

**Phone frame error count**
Ref: 
[ ky ][ ou ][ n ][ o ][ h ][ o ][ q ][ k ][ a ] ....

Hyp: 
[ k ][ i ][ n ][ ou ][ n ][ o ][ t ][ ou ][ k ][ a ] .... (Error: 35)
Minimum Phone Error (Povey 2002); Decision boundaries

MPE loss function:
\[ f^{MPE}_\Lambda = \sum_r \sum_n \mathcal{E}_{n,r} P_\Lambda (S_n | \mathcal{X}_r) \]

“posterior”:
\[ P_\Lambda (S_n | \mathcal{X}_r) = \frac{P(S_n \psi \eta p_\Lambda (\mathcal{X}_r | S_n) \psi}{\sum_k P(S_k \psi \eta p_\Lambda (\mathcal{X}_r | S_k) \psi} \]

\[ P_\Lambda (S_n | \mathcal{X}_r) = \frac{1}{1 + \left( \sum_{k \neq n} \exp \left( \log \left( P(S_k \psi \eta p_\Lambda (\mathcal{X}_r | S_k) \psi) \right) \right) \right)} \]

\[ m_{k,n} (\mathcal{X}_r, \Lambda) = \log \left( P(S_k \eta p_\Lambda (\mathcal{X}_r | S_k) \right) - \log \left( P(S_n \eta p_\Lambda (\mathcal{X}_r | S_n) \right) \]

pair-wise comparisons

\[ P_\Lambda (S_n | \mathcal{X}_r) = \frac{1}{1 + \sum_{k \neq n} \exp \left( \psi m_{k,n} (\mathcal{X}_r, \Lambda) \right) \]
MPE as multi-dimensional sigmoid

$$F_{MPE} = \sum_n \frac{\mathcal{E}_{n,r}}{1 + \sum_{k \neq n} \exp (\psi m_{k,n} (x_r, \Lambda))}$$
MPE derivative - String picture

\[
\frac{\partial f_{\Lambda}^{\text{MPE}}(\mathcal{X}_r)}{\partial \log p_{\Lambda}(\mathcal{X}_r, S_i)} = P_{\Lambda}(S_i | \mathcal{X}_r) \left( \mathcal{E}_{i,r} - f_{\Lambda}^{\text{MPE}}(\mathcal{X}_r) \right)
\]

Posterior * (Error – Average Error)

< 0  \rightarrow \text{“Correct”} \rightarrow \text{should improve i–th hyp}

> 0  \rightarrow \text{“Incorrect”} \rightarrow \text{should worsen i–th hyp}
Modified Forward-Backward for MPE over lattices

Forward probability mass for arc \( q \):

\[
\alpha_q = \sum_i \alpha_i p_i
\]

Forward average error for arc \( q \):

\[
\alpha'_q = \frac{\sum_i \alpha_i \alpha'_i}{\sum_i \alpha_i} + e_q
\]

Reference:
「今日の北海道の天気は晴れです」

“Today’s Hokkaido weather: sunny”
Overview

Background

Unification of Margin-modified MPE and MMI

Large-scale discriminative training
Minimum Phone Error
Margin-based variants; bMMI

MPE derivative - Arc picture

\[ \alpha'_q = \frac{\sum_i \alpha_i \alpha'_i}{\sum_i \alpha_i} + e_q \]

"Today’s"

"Yesterday’s"

... Average error of all strings through arc q

\[ e_q = 1 \]

Local error for arc q

\[ c_q = \alpha'_q + \beta'_q \]

\[ \gamma(q) = P_{\Lambda}(q|X_r) = \frac{\alpha_q \beta_q}{\alpha_{\text{final}}} = \frac{\alpha_q \beta_q}{\beta_{\text{start}}} \]

Arc Occupancy

\[ \frac{\partial F_{MPE}(X_r, \Lambda)}{\partial \log(p_{\Lambda}(q, X_r))} = P_{\Lambda}(q|X_r)(c_q - F_{MPE}(X_r, \Lambda)) \]
New approaches based on margin

- Intuition: improve generalization by making the training problem “harder”.
- “Large-margin MCE” (Yu et al., 2007)
  - Extension of McDermott & Katagiri (2004)’s Parzen window analysis of MCE → iteratively increase MCE sigmoid bias term
- Applicable to *implicit error models*:
  - “Large-margin HMMs” (Sha & Saul, 2007): Insertion of fine-grained error (e.g. Edit Distance) into the margin term
  - “Boosted MMI” (Povey et al., Saon & Povey, 2008)
- Heigold’s unified theory (2008): bring margin to standard MMI/MPE/MCE approaches
Linking ASR and Machine Learning

Machine learning community

- VC dimension
- SVM
- Large margin

ASR community

- MMI/MCE
- MPE
- “boosted” MMI

Heigold (ICML 2008, Interspeech 2008)
Modifying MPE/MMI with margin term

“Boost” likelihoods (Povey & Saon (2008), Heigold (2008)): 

\[ P(S_k)^\eta p_\Lambda(\mathcal{X}_r|S_k) \rightarrow P(S_k)^\eta p_\Lambda(\mathcal{X}_r|S_k)e^{\sigma \mathcal{E}_{k,r}} \]

\[ f_{\Lambda,\sigma}^{\text{MPE}} = \sum_r \frac{\sum_n E_{n,r} P(S_n)^\psi p_\Lambda(\mathcal{X}_r|S_n)^\psi e^{\psi \sigma \mathcal{E}_{n,r}}}{\sum_k P(S_k)^\psi p_\Lambda(\mathcal{X}_r|S_k)^\psi e^{\psi \sigma \mathcal{E}_{k,r}}} \]
Margin-modified MPE cost function:

\[ f^{MPE}_{\Lambda, \sigma} = \sum_r \frac{\sum_n \mathcal{E}_{n,r} P(S_n)^{\psi \eta} p_{\Lambda}(X_r | S_n)^{\psi} e^{\psi \sigma \mathcal{E}_{n,r}}}{\sum_k P(S_k)^{\psi \eta} p_{\Lambda}(X_r | S_k)^{\psi} e^{\psi \sigma \mathcal{E}_{k,r}}} \]

Re-write MPE cost function using comparison variables

\[ m_{k,n}(X_r, \Lambda) = \log(P(S_k)^{\eta} p_{\Lambda}(X_r | S_k)) - \log(P(S_n)^{\eta} p_{\Lambda}(X_r | S_n)) \]

\[ \Delta_{r}^{k,n} = \mathcal{E}_{k,r} - \mathcal{E}_{n,r} \quad \text{(error differential)} \]

\[ f^{MPE}_{\Lambda, \sigma} = \sum_r \sum_n \frac{\mathcal{E}_{n,r}}{1 + \sum_{k \neq n} e^{\psi (m_{k,n}(X_r, \Lambda) + \sigma \Delta_{r}^{k,n})}} \]
Effect of margin on MPE loss

\[ m_{1,0} \leq -\frac{\sigma}{\psi} (\mathcal{E}_{1,r} - \mathcal{E}_{0,r}) \]

\[ m_{2,0} \leq 0 \quad (\mathcal{E}_{2,r} = \mathcal{E}_{0,r}) \]

Shift by

\[ \frac{\sigma}{\psi} (\mathcal{E}_{1,r} - \mathcal{E}_{0,r}) \]
Margin-modified MMI (Povey & Saon, 2008)

Margin-modified MMI cost function:

\[
F_{\Lambda, \sigma}^{\text{MMI}} = -\frac{1}{\psi} \sum_r^R \log \frac{P(S_r)^{\psi \eta} p_{\Lambda}(x_r | S_r)^{\psi}}{\sum_k P(S_k)^{\psi \eta} p_{\Lambda}(x_r | S_k)^{\psi} e^{\psi \sigma \xi_k, r}}
\]

Re-write MMI cost function using comparison variables

\[
m_{k, n}(x_r, \Lambda) = \log (P(S_k)^{\eta} p_{\Lambda}(x_r | S_k)) - \log (P(S_n)^{\eta} p_{\Lambda}(x_r | S_n))
\]

\[
F_{\Lambda, \sigma}^{\text{MMI}} = \frac{1}{\psi} \sum_r^R \log \left( 1 + \sum_{k \neq r} e^{\psi (m_{k, r}(x_r, \Lambda) + \sigma \xi_k, r)} \right)
\]

MMI is a smooth approximation of the hinge loss(?!)}
Effect of margin on MMI loss

Shift by $\sigma \varepsilon_{k,r}$

$\sigma \varepsilon_{j,r}$

$m_{k,r}$

$m_{j,r}$
2300h Arabic Broadcast News (GALE)
2000h English conversational telephone speech (CTS)
“Boost” likelihoods (Povey & Saon (2008), Heigold (2008)):

\[
P(S_k)^{\eta} p_\Lambda (\chi_r | S_k) \rightarrow P(S_k)^{\eta} p_\Lambda (\chi_r | S_k) e^{\sigma \epsilon_{k,r}}
\]

\[
f_{MPE}^{\Lambda, \sigma} = \sum_r \frac{\sum_n \epsilon_{n,r} P(S_n)^{\psi} p_\Lambda (\chi_r | S_n)^{\psi} e^{\psi \sigma \epsilon_{n,r}}}{\sum_k P(S_k)^{\psi} p_\Lambda (\chi_r | S_k)^{\psi} e^{\psi \sigma \epsilon_{k,r}}}
\]

\[
F_{MMI}^{\Lambda, \sigma} = -\frac{1}{\psi} \sum_r \frac{P(S_r)^{\psi} p_\Lambda (\chi_r | S_r)^{\psi}}{\sum_k P(S_k)^{\psi} p_\Lambda (\chi_r | S_k)^{\psi} e^{\psi \sigma \epsilon_{k,r}}}
\]
dMMI: the “integrated” framework

Margin-space integration of MPE loss via differencing of MMI functionals for generalized error-weighted discriminative training

McDermott & Nakamura, Interspeech 2009

- Mathematical link between margin-modified MPE and MMI;
- Proposal of “dMMI”
MPE is the derivative of modified MMI!

\[
\frac{\partial}{\partial \sigma} F_{\Lambda,\sigma}^{MMI} = \frac{1}{\psi} \sum_r^R \frac{\partial}{\partial \sigma} \log \left( \sum_k P(S_k)^{\psi \eta} p_\Lambda (X_r | S_k)^{\psi} e^{\psi \sigma \mathcal{E}_{k,r}} \right)
\]

\[
= \sum_r^R \sum_n P(S_n)^{\psi \eta} p_\Lambda (X_r | S_n)^{\psi} e^{\psi \sigma \mathcal{E}_{n,r} \mathcal{E}_{n,r}}
\]

\[
= \sum_r \frac{\sum_k P(S_k)^{\psi \eta} p_\Lambda (X_r | S_k)^{\psi} e^{\psi \sigma \mathcal{E}_{k,r}}}{\sum_k P(S_k)^{\psi \eta} p_\Lambda (X_r | S_k)^{\psi} e^{\psi \sigma \mathcal{E}_{k,r}}}
\]

\[
= f_{\Lambda,\sigma}^{MPE}
\]
dMMI definition

Using previous result & Fundamental Theorem of Calculus:

\[
\begin{align*}
\int_{\sigma_1}^{\sigma_2} f_{\Lambda, \sigma}^{\text{MPE}} \, d\sigma &= \frac{F_{\Lambda, \sigma_2}^{\text{MMI}} - F_{\Lambda, \sigma_1}^{\text{MMI}}}{\sigma_2 - \sigma_1}
\end{align*}
\]
dMMI in practice

Just use “reverse-boosted” denominator lattice as numerator lattice:

\[
\mathcal{f}_{\Lambda, \sigma_1, \sigma_2}^{dMMI} = \frac{1}{\psi (\sigma_2 - \sigma_1)} \sum_r \log \frac{\sum_k P(S_k)^\psi \eta p_\Lambda (x_r | S_k)^\psi e^{\psi \sigma_2 \epsilon_{k,r}}}{\sum_k P(S_k)^\psi \eta p_\Lambda (x_r | S_k)^\psi e^{\psi \sigma_1 \epsilon_{k,r}}}
\]
Approximating MPE

As margin interval is reduced, dMMI converges to MPE

\[
\lim_{\sigma_1, \sigma_2 \to \sigma_c} f_{\Lambda, \sigma_1, \sigma_2}^{dMMI} = f_{\Lambda, \sigma_c}^{MPE}
\]

Property must hold for any correct implementations of bMMI and MPE!
Integrated view of discriminative training

\[ f(\sigma) \]

Weighted Error (e.g. MPE)

```
\text{``boosted''}
\text{MMI} = F(-\infty, \sigma_2)
```

\[ f_{MPE}(0) \]

\text{standard MPE}

```
F(0, \sigma_2)
```

\text{``boosted''}
\text{MPE}

\text{standard MPE}

\[ \sigma = 0 \quad \sigma = \sigma_2 > 0 \]

Margin-Integrated Weighted Error

\[ (= \text{Generalized MMI}) \]

\[ F(\sigma_1, \sigma_2) = \int_{\sigma_1}^{\sigma_2} f_{MPE}(\sigma) d\sigma \]
Leveraging approximated, shifted hinge functions
Gradient-based optimization using dMMI

MMI derivative w.r.t. given arc \( q \) in recognition lattice:

\[
\frac{\partial F^{M_MI}_{\Lambda,\sigma}}{\partial \log p^{\Lambda}_{\sigma} (q, \mathcal{X}_r)} = \gamma_{q,\Lambda,\sigma, \text{occupancy}}
\]

dMMI: run Forward-Backward twice, once for each margin:

\[
\frac{\partial f^{dM_MI}_{\Lambda,\sigma_1,\sigma_2}}{\partial \log p^{\Lambda}_{\sigma} (q, \mathcal{X}_r)} = \frac{\gamma_{q,\Lambda,\sigma_2} - \gamma_{q,\Lambda,\sigma_1}}{\sigma_2 - \sigma_1} = \gamma_{q,\Lambda,\sigma_1,\sigma_2}^{dM_MI}
\]

Overall dMMI gradient:

\[
\frac{\partial f^{dM_MI}_{\Lambda,\sigma_1,\sigma_2}}{\partial \Lambda_i} = \sum_r \sum_q \gamma_{q,\Lambda,\sigma_1,\sigma_2}^{dM_MI} \frac{\partial \log p^{\Lambda}_{\sigma} (q, \mathcal{X}_r)}{\partial \Lambda_i}
\]
dMMI as integral over margin prior

\[
f_{dMMI}^\Lambda,p(\sigma) = \int_{-\infty}^{\infty} f_{MPE}^\Lambda,\sigma p(\sigma) \, d\sigma
\]

\[
p(\sigma) = \begin{cases} 
0 & \sigma < \sigma_1 \\
\frac{1}{\sigma_2 - \sigma_1} & \sigma_1 \leq \sigma \leq \sigma_2 \\
0 & \sigma > \sigma_2
\end{cases}
\]
dMMI as building block for modeling general margin priors

\[
f_{d\text{MMI}}_{\Lambda,p(\sigma)} = \int_{-\infty}^{\infty} f_{\Lambda,\sigma}^{\text{MPE}} p(\sigma) d\sigma \\
\approx \frac{c}{N} \sum_{n=1}^{N} \int_{\sigma_1(n)}^{\sigma_2(n)} f_{\Lambda,\sigma}^{\text{MPE}} d\sigma = \frac{c}{N} \sum_{n=1}^{N} \left( \mathcal{F}_{\Lambda,\sigma_2(n)}^{\text{MMI}} - \mathcal{F}_{\Lambda,\sigma_1(n)}^{\text{MMI}} \right)
\]
Numerical approximation of arbitrary margin priors

- E.g. prior $p(\sigma) = c \exp(-c|\sigma|)$ used for Minimum Relative Entropy Discrimination, Jebara (2004)
- Here: use prior in context of standard HMM-based discriminative training
- Approximate prior using sum of step functions (cf Lebesgue integration)
Building margin prior using dMMI

Desired margin prior:
\[ p(\sigma) = c \exp(-c|\sigma|) \]

1. Divide \( y = [0, c] \) into \( N+1 \) equal parts
2. Find left & right inverses
3. Sum step functions to approximate prior

\[ p_n(\sigma) = \begin{cases} 
0 & : \sigma < \sigma_1(n) \\
\frac{c}{N} & : \sigma_1(n) \leq \sigma \leq \sigma_2(n) \\
0 & : \sigma > \sigma_2(n) 
\end{cases} \]

\[ p(\sigma) \approx \sum_{n=1}^{N} p_n(\sigma) \]

\[ \sigma_1(n) = -p^{-1}(y(n)) = \frac{1}{c} \log \left( \frac{n}{N+1} \right) \]

\[ \sigma_2(n) = p^{-1}(y(n)) = -\frac{1}{c} \log \left( \frac{n}{N+1} \right) \]
Summary

- MPE explicitly models non-uniform error, e.g. phone or word error including insertions, deletions & substitutions
- Margin-based “Boosted MMI” (bMMI):
  - super-cheap approach for incorporating non-uniform error into loss function;
  - however objective is still (modified) Mutual Information, not explicit model of error.
- “Differenced MMI” (dMMI) is similarly cheap alternative that
  - is explicitly linked to error;
  - generalizes MPE;
  - possibly offers better performance (Delcroix et al. ICASSP 2012; Kubo et al. Interspeech 2012);
  - can be further generalized to define arbitrary margin priors for lattice-based discriminative training.