# MMI-MAP and MPE-MAP for Acoustic Model Adaptation

Dan Povey, Mark Gales, Do Yeong Kim & Phil Woodland

Sep 3 2003

Cambridge University Engineering Department

# Overview

- Maximum A Posteriori (MAP) is a standard adaptation scheme:
  - increasing adaptation data tends to Maximum Likelihood estimation;
  - referred to as `ML-MAP` is this talk.
- This paper describes two new discriminative MAP schemes:
  - increasing adaptation data tends to discriminative estimation;
  - maximum mutual information (`MMI-MAP`) and minimum phone error (`MPE-MAP`) adaptation investigated.
- Two applications will be described:
  - task port: from SwitchBoard to VoiceMail;
  - gender dependent models: GD models for Broadcast News.

# Discriminative Training Criteria

- The discriminative criteria considered are:
  - Maximum mutual information (MMI)

$$\mathcal{F}^{\mathrm{MMI}}(\lambda) = \log \frac{p_{\lambda}\left(\mathcal{O}|\mathcal{M}_s\right)^{\kappa} P(s)^{\kappa}}{\sum_s p_{\lambda}\left(\mathcal{O}_r|\mathcal{M}_s\right)^{\kappa} P(s)^{\kappa}}$$

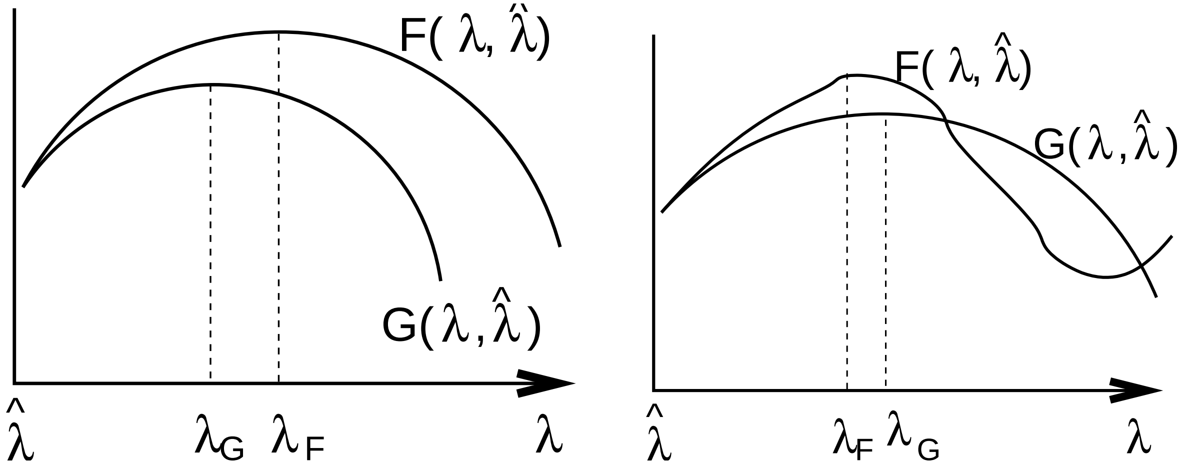
  - Minimum Phone Error (MPE)

$$\mathcal{F}^{\mathrm{MPE}}(\lambda) = \frac{\sum_s p_{\lambda}(\mathcal{O}|\mathcal{M}_s)^{\kappa} P(s)^{\kappa} \mathrm{RawAccuracy}(s)}{\sum_s p_{\lambda}(\mathcal{O}|\mathcal{M}_s)^{\kappa} P(s)^{\kappa}}$$

    $\mathrm{RawAccuracy}(s)$ is a measure of the number of phones accurately transcribed.

- An alternative perspective on discriminative parameter estimation is described.
- Discriminative MAP schemes within this framework will be described.

# Strong/Weak Sense Auxiliary Functions




(a)Strong Sense (b) Weak Sense

- Strong Sense: used for standard EM - guaranteed convergence, requires

$$\mathcal{G}(\lambda, \hat{\lambda}) - \mathcal{G}(\hat{\lambda}, \hat{\lambda}) \leq \mathcal{F}(\lambda) - \mathcal{F}(\hat{\lambda}),$$

- Weak Sense: applicable to MMI - yields Extended BW, requires

$$\frac{\partial}{\partial\lambda}\mathcal{G}(\lambda, \hat{\lambda})\bigg|_{\lambda=\hat{\lambda}} = \frac{\partial}{\partial\lambda}\mathcal{F}(\lambda)\bigg|_{\lambda=\hat{\lambda}}.$$

# Weak Sense Auxiliary functions for MMI

- MMI criterion may be expressed as (ignoring $\kappa$ for simplicity)

$$\mathcal{F}^{\mathrm{MMI}}(\lambda) = \log p(\mathcal{O}|\mathcal{M}^{\mathrm{num}}) - \log p(\mathcal{O}|\mathcal{M}^{\mathrm{den}})$$

- The weak sense auxiliary function is

$$\mathcal{G}^{\mathrm{MMI}}(\lambda, \hat{\lambda}) = \mathcal{G}^{\mathrm{num}}(\lambda, \hat{\lambda}) - \mathcal{G}^{\mathrm{den}}(\lambda, \hat{\lambda}) + \mathcal{G}^{\mathrm{sm}}(\lambda, \hat{\lambda}).$$

where $\mathcal{G}^{\mathrm{num}}(\lambda, \hat{\lambda})$ and $\mathcal{G}^{\mathrm{den}}(\lambda, \hat{\lambda})$ are standard strong sense auxiliary functions.

- A smoothing term is added to improve stability - satisfies

$$\left.\frac{\partial}{\partial \lambda}\mathcal{G}^{\mathrm{sm}}(\lambda, \hat{\lambda})\right|_{\lambda=\hat{\lambda}} = 0$$

This ensures that final function is still a valid weak sense auxiliary function

# MMI Updates

- A possible smoothing function is

$$
\begin{aligned}
\mathcal{G}^{\mathrm{sm}}(\lambda, \hat{\lambda}) &= \sum_{j=1}^{J} -D_j \frac{1}{2}\left( \log(2\pi\sigma_j^2) + \frac{(\hat{\mu}_j^2 + \hat{\sigma}_j^2) - 2\hat{\mu}_j\mu_j + \mu_j^2}{\sigma_j^2} \right) \\
&= \sum_{j=1}^{J} \mathcal{Q}(D_j, D_j\hat{\mu}_j, D_j(\hat{\mu}_j^2 + \hat{\sigma}_j^2), \lambda_j)
\end{aligned}
$$

- This yields the following MMI update for the means

$$
\mu_j = \frac{\left\{\theta_j^{\mathrm{num}}(\mathcal{O}) - \theta_j^{\mathrm{den}}(\mathcal{O})\right\} + D_j\hat{\mu}_j}{\left\{\gamma_j^{\mathrm{num}} - \gamma_j^{\mathrm{den}}\right\} + D_j}
$$

Same as the standard Extended Baum-Welch update formulae.

# Incorporating Prior Information

- By definition a function is a weak sense auxiliary function of itself:
  - a log-prior may be directly added to the weak sense auxiliary function.

- Consider using the ML estimate as the centre prior

$$\log p(\lambda_j) = \mathcal{Q}\left(\tau^I, \tau^I \mu_j^{\text{ml}}, \tau^I(\mu_j^{\text{ml2}} + \sigma_j^{\text{ml2}}), \lambda_j\right)$$

where $\mu_j^{\text{ml}} = \frac{\theta_j^{\text{num}}(\mathcal{O})}{\gamma_j^{\text{num}}}$.

- This yields I-Smoothing

$$\mu_j = \frac{\{\theta_j^{\text{num}}(\mathcal{O}) - \theta_j^{\text{den}}(\mathcal{O})\} + D_j\hat{\mu}_j + \tau^I\mu_j^{\text{ml}}}{\{\gamma_j^{\text{num}} - \gamma_j^{\text{den}}\} + D_j + \tau^I}$$

  - $\tau^I$ determines influence of "prior" (ML estimate) on the final MMI estimate.

# MMI-MAP

- For adaptation/porting the ML estimate may not be robust
  - use a ML-MAP estimate as the prior

- Use count-smoothing ML-MAP with prior parameters $(\tilde{\mu}_j)$

$$\mu_j = \frac{\{\theta_j^{\text{num}}(\mathcal{O}) - \theta_j^{\text{den}}(\mathcal{O})\} + D_j\hat{\mu}_j + \tau^I\mu_j^{\text{map}}}{\{\gamma_j^{\text{num}} - \gamma_j^{\text{den}}\} + D_j + \tau^I}$$

where $\mu_j^{\text{map}} = \frac{\theta_j^{\text{num}}(\mathcal{O})+\tau\tilde{\mu}_j}{\gamma_j^{\text{num}}+\tau}$

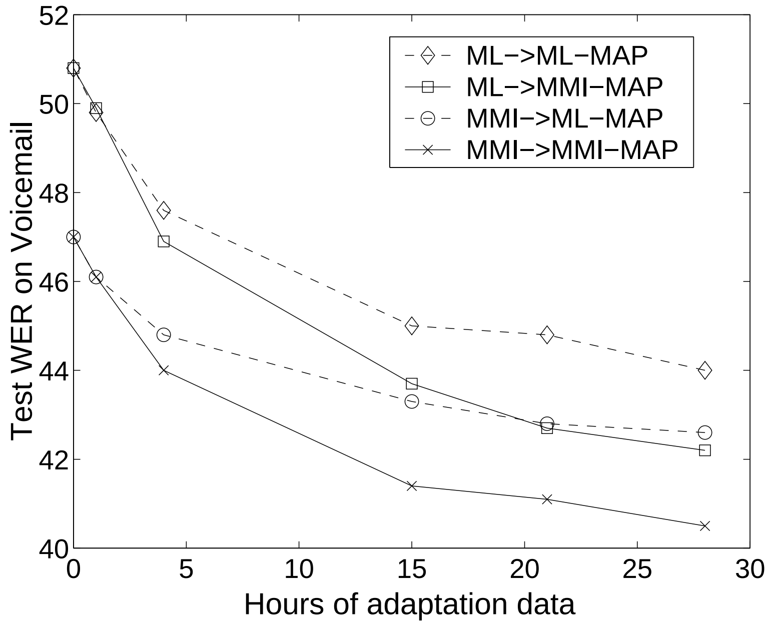
- Two smoothing variables for MMI-MAP
  - $\tau$ determines how "close" the prior is to the ML estimate
  - $\tau^I$ determines how much the prior influences the final estimate.

- Similar form may be used for MPE-MAP.

# Switchboard to VoiceMail Porting

- Switchboard (source) - spontaneous telephone speech task
  - 265 hours of training data, state-of-the-art system;
  - gender-independent cross-word state clustered triphones;
  - 6684 distinct states, 16 components per state;
  - Systems trained using ML and MMI training.
- VoiceMail (target) - VoiceMail message data:
  - voicemail messages collected by IBM employees;
  - 28 hours of acoustic data (partitioned into 5 sets);
  - 1.5 hour test set (1 hour taken from second release of training data).
- Standard Switchboard evaluation language model used.

# Switchboard to Voicemail Porting Results




- WERs on Voicemail for varying amounts of adaptation data
- (MMI or ML) adapted with (MMI-MAP or ML-MAP)
- 4.5% relative improvement from MMI-MAP vs. ML-MAP (starting from MMI) @ 30h adaptation data

# Gender Adaptation on Broadcast News

- 142 hours of training data (`BNtrain97` and `BNtrain98`)
- Cross-word state clustered triphones;
- 6,976 distinct states, 16 components per state;
- Standard front end (Std), MF-PLP plus first and second-order deltas;
- Heteroscedastic linear discriminant analysis (HLDA):
  - expand feature vector using third-order deltas;
  - linear projection back to 39 dimensions.
- `BNeval98` test set.

# BN Gender Adaptation Results

| System | WER (%) | |
|---|---|---|
| | Std | HLDA |
| MLE-GI | 19.6 | 17.9 |
| MLE-GD | 18.8 | 17.1 |
| MMI-GI | 17.0 | — |
| MPE-GI | 16.2 | 15.0 |
| $\rightarrow$MPE-MAP | 15.7 | 14.5 |

- With ML system, gender adaptation (using ML-MAP) gave 0.8% absolute
- With MPE system, MPE-MAP gave 0.5% absolute
- MPE+MPE-MAP system 14.5% WER, vs. 17.1% for MLE (Both with HLDA)
- MPE with GD training gives 14.8%

# Summary

- Extended the MAP adaptation technique to be used with discriminative training: both MMI and MPE

- Tends to discriminative training performance with infinite adaptation data

- Increases effectiveness of MAP with discriminatively-trained models

- Improvements over MLE-MAP for
  - Task adaptation for Switchboard$\rightarrow$ Voicemail
  - Creation of Broadcast News gender-dependent models