# DISCRIMINATIVE MAP FOR ACOUSTIC MODEL ADAPTATION

*D. Povey, P.C. Woodland & M.J.F. Gales*

Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {`dp10006,pcw,mjfg`}`@eng.cam.ac.uk`

**ABSTRACT**

In this paper we show how a discriminative objective function such as Maximum Mutual Information (MMI) can be combined with a prior distribution over the HMM parameters to give a discriminative Maximum A Posteriori (MAP) estimate for HMM training. The prior distribution can be based around the Maximum Likelihood (ML) parameter estimates, leading to a technique previously referred to as I-smoothing; or for adaptation it can be based around a MAP estimate of the ML parameters, leading to what we call MMI-MAP. This latter approach is shown to be effective for task adaptation, where data from one task (Voicemail) is used to adapt a HMM set trained on another task (Switchboard). It is shown that MMI-MAP results in a 2.1% absolute reduction in word error rate relative to standard ML-MAP with 30 hours of Voicemail task adaptation data starting from a MMI-trained Switchboard system.

## 1. INTRODUCTION

Recently discriminative training techniques such as Maximum Mutual Information Estimation (MMIE) have been shown to outperform conventional Maximum Likelihood Estimation (MLE) for large vocabulary HMM-based speech recognition [8]. However adaptation techniques for these models such are still generally based on MLE: for intance, Maximum Likelihood Linear Regression (MLLR) and Maximum A Postriori (MAP) adaptation.

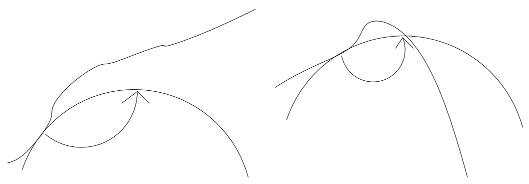
While it has been shown that MLLR can be effective for speaker adaptation of MMI-trained models [8] and conventional MAP can be effective for task adaptation of MMI-trained models [1] it is interesting to investigate what additional benefits can be found for using discriminative objective functions in adaptation as well as for training the original models. Previous work in discrimative adaptation includes a MAP-type scheme described in [3] and discriminative transform estimation [7].

This paper introduces a technique denoted MMI-MAP which combines a prior distribution with the statistics required for MMI estimation using the adaptation data. A key feature of the MMI-MAP scheme presented here is that it is a two-level scheme with the prior derived by conventional (ML)-MAP estimation. The technique is evaluated on task adaptation, adapting initial HMM sets trained on the Switchboard system to the Voicemail task.

The paper is arranged as follows. In Section 2 the concept of weak-sense auxiliary functions is introduced, which is a convenient framework for deriving discriminative parameter updates. This is used to derive the update formulae for MMI. Section 3 describes the use of weak-sense auxiliary functions for the case of a prior distribution and introduces MMI-MAP. Section 4 presents the experimental results on task adaptation.

This work was funded by the European Commission under the Language project Le-5 Coretex. Extensive use was made of equipment donated by IBM under an SUR award.

## 2. STRONG-SENSE AND WEAK-SENSE AUXILIARY FUNCTIONS



(a) (b)

**Fig. 1**. Use of (a) strong-sense and (b) weak-sense auxiliary functions for function optimisation

If a function $\mathcal{F}(\lambda)$ is to be maximised, then $\mathcal{G}(\lambda, \lambda')$ is a *strong-sense auxiliary function* for $\mathcal{F}(\lambda)$ *around* $\lambda'$, iff

$$\mathcal{G}(\lambda, \lambda') - \mathcal{G}(\lambda', \lambda') \leq \mathcal{F}(\lambda) - \mathcal{F}(\lambda'), \tag{1}$$

where $\mathcal{G}(\lambda, \lambda')$ is a smooth function of $\lambda$. A strong-sense auxiliary function is the kind of auxiliary function used in Expectation-Maximisation (EM). The idea is illustrated in Figure 1(a). A maximum w.r.t. $\lambda$ of the function $\mathcal{G}(\lambda, \lambda')$ is found, indicated by the arrow. If this increases $\mathcal{G}$, it will also increase $\mathcal{F}$; if $\mathcal{G}$ is at a local maximum then $\mathcal{F}$ is also at a local maximum. These conditions follow from Eq. (1), and imply that repeated maximisation of the auxiliary function is guaranteed to converge to a local maximum of $\mathcal{F}(\lambda)$.

A *weak-sense auxiliary function* for $\mathcal{F}(\lambda)$ *around* $\lambda'$ is a smooth function $\mathcal{G}(\lambda, \lambda')$ such that

$$\left.\frac{\partial}{\partial \lambda}\mathcal{G}(\lambda, \lambda')\right|_{\lambda=\lambda'} = \left.\frac{\partial}{\partial \lambda}\mathcal{F}(\lambda)\right|_{\lambda=\lambda'}. \tag{2}$$

The idea is shown in Figure 1(b). The gradients of the two functions are the same around the point $\lambda = \lambda'$. Maximising the function $\mathcal{G}(\lambda, \lambda')$ w.r.t. $\lambda$ does not now guarantee an increase in $\mathcal{F}(\lambda)$. However if there is no change in $\lambda$ after maximisation on a particular iteration, this implies that we have reached a local maximum of $\mathcal{F}(\lambda)$ (the gradient is zero at that point). If the update converges it will be to a local maximum of $\mathcal{F}(\lambda)$.

The condition of being a weak-sense auxiliary function can be considered a minimum condition for an auxiliary functon used for optimisation. In addition the function should be chosen so as to ensure good convergence.

Weak-sense auxiliary functions are useful when optimising functions containing some terms that can be optimised by strong-sense auxiliary functions but others that cannot; one such function is the MMI objective function (expressed as a sum of logs). Weak-sense auxiliary functions make it possible to modify the update procedures used in techniques based on Expectation-Maximisation (i.e, the use of strong-sense auxiliary functions), rather than using entirely different techniques based on gradient descent.

### 2.1. Strong-sense auxiliary functions for ML estimation

An HMM likelihood (for an observation sequence) is a sum over state sequences. In general terms, if the HMM parameters are represented by $\lambda$, the HMM likelihood $\mathcal{F}(\lambda) = \log \sum_x f_x(\lambda)$ is maximised, where $x$ correspond to state sequences and $f_x(\lambda)$ are state sequence likelihoods. If the optimisation is started at $\lambda = \lambda'$, a strong-sense auxiliary function for $\mathcal{F}(\lambda)$ is

$$\mathcal{G}(\lambda, \lambda') = \sum_x \frac{f_x(\lambda')}{\sum_y f_y(\lambda')} \log f_x(\lambda). \tag{3}$$

Eq. (1) can be shown to hold for the $\mathcal{F}(\lambda)$ and $\mathcal{G}(\lambda, \lambda')$ of Eq. (3); it reduces to an equation involving the Kullback-Leibler distance.

The auxiliary function is a sum of state-sequence log likelihoods $\log f_x(\lambda)$, weighted by the initial posterior probability $\frac{f_x(\lambda')}{\sum_y f_y(\lambda')}$ of the state sequence.

If $\gamma_j(t)$ is defined as the sum over state sequence posterior probabilities $\frac{f_x(\lambda')}{\sum_y f_y(\lambda')}$ for all sequences $x$ that include state $j$ at time $t$, the auxiliary function is as follows (considering only the Gaussian parameters and a single Gaussian per state):

$$\mathcal{G}(\lambda, \lambda') = \sum_{j=1}^{J} \sum_{t=1}^{T} \gamma_j(t) \log \mathcal{N}(\mathcal{O}(t)|\mu_j, \sigma_j^2). \tag{4}$$

where $\mu_j$ and $\sigma_j^2$ are the updated mean and variance corresponding to the new parameters $\lambda$. This can equivalently be expressed as

$$\begin{aligned}\mathcal{G}(\lambda, \lambda') &= \sum_{j=1}^{J} -\frac{1}{2}\left(\gamma_j \log(2\pi\sigma^2) + \frac{\theta_j(\mathcal{O}^2) - 2\theta_j(\mathcal{O})\mu_j + \gamma_j\mu_j^2}{\sigma_j^2}\right) \\ &= \sum_{j=1}^{J} Q(\gamma_j, \theta_j(\mathcal{O}), \theta_j(\mathcal{O}^2)|\mu_j, \sigma_j^2)\end{aligned} \tag{5}$$

where $\theta_j(\mathcal{O}) = \sum_{t=1}^{T} \gamma_j(t)\mathcal{O}(t)$ is the sum of data weighted by probability, $\theta_j(\mathcal{O}^2)$ is the same sum over squared data and $\gamma_j = \sum_{t=1}^{T} \gamma_j(t)$ is the occupancy of the state.

### 2.2. Weak-sense auxiliary functions for MMI estimation

The MMI objective function is a difference of HMM log likelihoods, $\mathcal{F}(\lambda) = \log P(\mathcal{O}|\mathcal{M}^{\text{num}}) - \log P(\mathcal{O}|\mathcal{M}^{\text{den}})$, where $\mathcal{M}^{\text{num}}$ and $\mathcal{M}^{\text{den}}$ are HMMs corresponding to the correct transcription and all passible transcriptions respectively. Strong-sense auxiliary functions as for ML estimation, $\mathcal{G}^{\text{num}}(\lambda, \lambda')$ and $\mathcal{G}^{\text{den}}(\lambda, \lambda')$ can be derived separately for the two log-likelihoods $\log P(\mathcal{O}|\mathcal{M}^{\text{num}})$ and $\log P(\mathcal{O}|\mathcal{M}^{\text{den}})$: the auxiliary functions differ only in the model topology used. A difficulty arises because the second term is negated in the MMI objective function; strong-sense auxiliary functions cannot be used when the problem is negated since the inequality of Eq. (1) will no longer hold. However weak-sense auxiliary functions do not suffer from this problem, and the difference $\mathcal{G}^{\text{num}}(\lambda, \lambda') - \mathcal{G}^{\text{den}}(\lambda, \lambda')$ is still a weak-sense auxiliary function for the MMI objective function.

In order to improve convergence we can add a smoothing function $\mathcal{G}^{\text{sm}}(\lambda, \lambda')$ which can in principle be any function with a zero differential w.r.t. $\lambda$ around the current estimate $\lambda = \lambda'$. This will not affect the local differential and the result will be a still be a weak-sense auxiliary function for the MMI objective function. This leads to the following auxiliary function:

$$\mathcal{G}(\lambda, \lambda') = \mathcal{G}^{\text{num}}(\lambda, \lambda') - \mathcal{G}^{\text{den}}(\lambda, \lambda') + \mathcal{G}^{\text{sm}}(\lambda, \lambda'). \tag{6}$$

One possible form for $\mathcal{G}^{\text{sm}}(\lambda, \lambda')$ is:

$$\mathcal{G}^{\text{sm}}(\lambda, \lambda') = \sum_{j=1}^{J} Q(D_j, D_j{\mu'}_j, D_j({\mu'}_j^2 + {\sigma'}_j^2)|\mu_j, \sigma_j^2), \tag{7}$$

which has a zero differential w.r.t. the parameters $\sigma_j^2$ and $\mu_j$ evaluated at the old values ${\sigma'}_j^2$ and ${\mu'}_j$, so the auxiliary function is still a weak-sense auxiliary function for the objective function around $\lambda'$. $D_j$ are positive smoothing constants for each state $j$ (or each Gaussian $j, m$ in the mixture-of-Gaussians case).

The total auxiliary function (considering only terms involving Gaussian parameters) now becomes:

$$\begin{aligned}\mathcal{G}(\lambda, \lambda') = \sum_{j=1}^{J} & Q(\gamma_j^{\text{num}}, \theta_j^{\text{num}}(\mathcal{O}), \theta_j^{\text{num}}(\mathcal{O}^2)|\mu_j, \sigma_j^2) \\ & -Q(\gamma_j^{\text{den}}, \theta_j^{\text{den}}(\mathcal{O}), \theta_j^{\text{den}}(\mathcal{O}^2)|\mu_j, \sigma_j^2) \\ & +Q(D_j, D_j{\mu'}_j, D_j({\mu'}_j^2 + {\sigma'}_j^2)|\mu_j, \sigma_j^2).\end{aligned} \tag{8}$$

The above analysis can be extended for Gaussian mixture likelihoods with Gaussian components $m = 1 \ldots M$. For multiple components, maximisation of the function in Eq. 8 leads to the Extended Baum-Welch (EB) update equations [8] as follows:

$$\mu_{jm} = \frac{\left\{\theta_{jm}^{\text{num}}(\mathcal{O}) - \theta_{jm}^{\text{den}}(\mathcal{O})\right\} + D_{jm}{\mu'}_{jm}}{\left\{\gamma_{jm}^{\text{num}} - \gamma_{jm}^{\text{den}}\right\} + D_{jm}} \tag{9}$$

$$\sigma_{jm}^2 = \frac{\left\{\theta_{jm}^{\text{num}}(\mathcal{O}^2) - \theta_{jm}^{\text{den}}(\mathcal{O}^2)\right\} + D_{jm}({\sigma'}_{jm}^2 + {\mu'}_{jm}^2)}{\left\{\gamma_{jm}^{\text{num}} - \gamma_{jm}^{\text{den}}\right\} + D_{jm}} - \mu_{jm}^2 \tag{10}$$

where $D_{jm}$ are set on a per-Gaussian level to the larger of i) twice the smallest value needed to ensure positive variances, or ii) $\gamma_{jm}^{\text{den}}$ times a further constant $E$, which is generally set to 1 or 2. Since it is based on weak-sense auxiliary functions, if this update converges to a particular value $\lambda$ of the HMM parameters then $\lambda$ must be a local maximum of the objective function. For large $E$ the update approaches gradient descent with parameter-specific learning rates proportional to $\frac{1}{E}$, so there should be some sufficiently small learning rate (i.e. sufficiently large $E$) for which the update converges.

## 3. MAP UPDATES

Any function is both a weak and strong-sense auxiliary function of itself around any point. Therefore if we add a log prior distribution $\log P(\lambda)$ to the MMI objective function making

$$\mathcal{F}(\lambda) = \log P(\mathcal{O}|\mathcal{M}^{\text{num}}) - \log P(\mathcal{O}|\mathcal{M}^{\text{den}}) + \log P(\lambda), \tag{11}$$

the extra term can simply be added to the auxiliary function leading to

$$\mathcal{G}(\lambda, \lambda') = \mathcal{G}^{\text{num}}(\lambda, \lambda') - \mathcal{G}^{\text{den}}(\lambda, \lambda') + \mathcal{G}^{\text{sm}}(\lambda, \lambda') + \log P(\lambda). \tag{12}$$

### 3.1. Priors over Gaussian parameters

The form of prior used over a mean $\mu$ and variance $\sigma^2$ is:

$$\log P(\mu, \sigma^2) = k + Q(\tau, \tau\mu_{\text{prior}}, \tau(\sigma^2_{\text{prior}} + \mu^2_{\text{prior}})|\mu, \sigma^2), \tag{13}$$

where $Q(\dots)$, defined in Eq. 5, is the likelihood of $\tau$ points of data with mean $\mu_{\text{prior}}$ and variance $\sigma^2_{\text{prior}}$, and $k$ is a normalisation term which can be ignored.

For the mean, this prior is a Gaussian with variance $\frac{\sigma^2}{\tau}$, i.e. $\frac{1}{\tau}$ times the variance of the distribution itself, as in conventional MAP [4]. For the variance, defining $S = (\mu - \mu_{\text{prior}})^2 + \sigma^2_{\text{prior}}$, matching the first and second-order terms of the Taylor expansion around the value $\sigma^2 = S$ shows that the distribution over $\sigma^2$ is locally equivalent to a Gaussian distribution with mean $S$ and variance $\frac{2S^2}{\tau}$. The prior over the variance differs from the standard approach to priors over variances [4], in that the prior over the variance has a slightly different mean. This formulation makes sense if our intuition about the prior is that the Gaussian parameters ought to give a high likelihood to data drawn from a particular distribution.

### 3.2. I-smoothing

I-smoothing for discriminative training [6] may be regarded as the use of a prior over the parameters of each Gaussian, with the prior being based on the ML statistics. The log prior likelihood is equal to $Q(\tau^I, \tau^I \frac{\theta^{\text{num}}_{jm}(\mathcal{O})}{\gamma^{\text{num}}_{jm}}, \tau^I \frac{\theta^{\text{num}}_{jm}(\mathcal{O}^2)}{\gamma^{\text{num}}_{jm}}|\mu_{jm}, \sigma^2_{jm})$, i.e. the likelihood of $\tau^I$ points of data with mean and variance equal to the numerator (correct model) mean and variance. This can be implemented by altering the numerator statistics as follows:

$$\gamma^{\text{num}\prime}_{jm} = \gamma^{\text{num}}_{jm} + \tau^I \tag{14}$$

$$\theta^{\text{num}}_{jm}(\mathcal{O})' = \theta^{\text{num}}_{jm}(\mathcal{O}) \frac{\gamma^{\text{num}}_{jm} + \tau^I}{\gamma^{\text{num}}_{jm}} \tag{15}$$

$$\theta^{\text{num}}_{jm}(\mathcal{O}^2)' = \theta^{\text{num}}_{jm}(\mathcal{O}^2) \frac{\gamma^{\text{num}}_{jm} + \tau^I}{\gamma^{\text{num}}_{jm}} \tag{16}$$

Typically $\tau^I$ is set to around 100 for MMI training.

### 3.3. MMI-MAP

In the context of adapting a HMM set, the use of ML statistics accumulated from the data as the center of the prior may be non-robust since there may not be enough data to estimate the ML Gaussian parameters. In this case it is preferable to estimate the the center of the prior in a conventional ML-MAP fashion. The technique denoted MMI-MAP is the use of ML-MAP estimates of the Gaussian parameters to estimate the center of a prior used to smooth the MMI-trained parameters.

In the first level of MAP the unadapted mean and variance $\mu^{\text{orig}}_{jm}$ and $\sigma^{\text{orig}}_{jm}$ are used as the prior, and the numerator (ML) statistics as the evidence. The parameters $\mu_{\text{prior}}$ and $\sigma_{\text{prior}}$ of the second-level prior are obtained by maximising the product of the prior $Q(\tau^{\text{MAP}}, \tau^{\text{MAP}}\mu^{\text{orig}}_{jm}, \tau^{\text{MAP}}(\sigma^{\text{orig}\,2}_{jm} + \mu^{\text{orig}\,2}_{jm})|\mu_{\text{prior}}, \sigma^2_{\text{prior}})$ and the evidence $Q(\gamma^{\text{num}}_{jm}, \theta^{\text{num}}_{jm}(\mathcal{O}), \theta^{\text{num}}_{jm}(\mathcal{O}^2)|\mu_{\text{prior}}, \sigma^2_{\text{prior}})$.

In the second level of MAP, the log prior is $Q(\tau^I, \tau^I\mu_{\text{prior}}, \tau^I(\sigma^2_{\text{prior}} + \mu^2_{\text{prior}})|\mu_{jm}, \sigma^2_{jm})$ and the evidence is the MMI criterion itself. The prior can be included in the EB re-estimation process by adding the three moments of the data $\tau^I$, $\tau^I\mu_{\text{prior}}$ and $\tau^I(\sigma^2_{\text{prior}} + \mu^2_{\text{prior}})$ to the numerator (num) statistics of the Gaussian in a modification of Equations 14 to 16, using:

$$\begin{aligned} \tau^I\mu_{\text{prior}} &= \tau^I \frac{\theta^{\text{num}}_{jm}(\mathcal{O}) + \tau^{\text{MAP}}\mu^{\text{orig}}_{jm}}{\gamma^{\text{num}}_{jm} + \tau^{\text{MAP}}} \\ \tau^I(\sigma^2_{\text{prior}} + \mu^2_{\text{prior}}) &= \tau^I \frac{\theta^{\text{num}}_{jm}(\mathcal{O}) + \tau^{\text{MAP}}(\sigma^{\text{orig}\,2}_{jm} + \mu^{\text{orig}\,2}_{jm})}{\gamma^{\text{num}}_{jm} + \tau^{\text{MAP}}}. \end{aligned}$$

## 4. EXPERIMENTS

This section describes the evaluation of MMI-MAP for the task adaptation from HMMs trained on the Switchboard database to the Voicemail (VM) task.

### 4.1. Experimental conditions

Initial Switchboard HMM training used 265 hours of data, and task adaptation and testing was on the Voicemail database [5, 2]. The total amount of VM training/adaptation data is about 30h, and subsets of approximately 1h, 4h, 15h and 20h were also used [2]. The VM test data was 94 minutes long and is described in detail in [2]. All HMM sets had 6684 tree-clustered states and 16 Gaussians per state.

All test set WERs reported here are from testing with unadapted Switchboard language models (LMs). The use of adapted LMs greatly improves error rates [2], but the relative improvements from discriminative adaptation are similar.

Three Switchboard-trained models were used for training: an MLE-trained model, and both MMIE and MPE-trained [6] models. The implementation used lattices generated using bigram LMs but including unigram LM probabilities, scaling of the likelihoods during forward-backward alignment by the inverse of the usual LM scale factor, and the "exact-match" form of forward-backward alignment [8].

MMI-MAP task adaptation was continued for four iterations and ML-MAP for one iteration; further iterations of ML-MAP were not found to be helpful. The smoothing constant $E$ for Gaussian updates (Section 2.2) was set to 2. Conventional MAP (here referred to as ML-MAP) was performed with $\tau = 10$, MMI-MAP with $\tau^I = 100$ and $\tau^{\text{MAP}} = 10$.

### 4.2. Results

Figure 2 shows the effect of adapting an ML or MMI-trained initial HMM set with ML-MAP or with MMI-MAP. The improve-

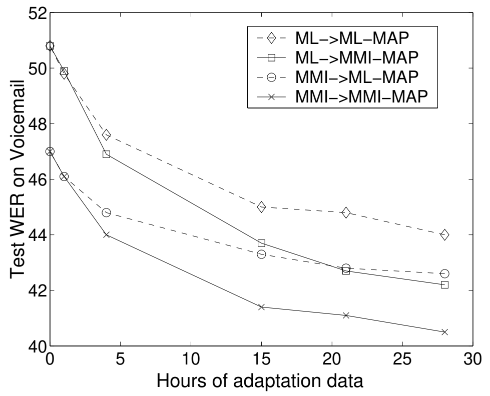



**Fig. 2**. MMI-MAP and ML-MAP from MMI and ML baselines: Unadapted LM

| Training, Adaptation: | Hours adaptation data 0 | 1 | 4 | 15 | 20 | 30 |
|---|---|---|---|---|---|---|
| ML,ML-MAP | 50.8 | 49.8 | 47.6 | 45.0 | 44.8 | 44.0 |
| ML,MMI-MAP | 50.8 | 49.9 | 46.9 | 43.7 | 42.7 | 42.2 |
| MMI,ML-MAP | 47.0 | 46.1 | 44.8 | 43.3 | 42.8 | 42.6 |
| MMI,MMI-MAP | 47.0 | 46.1 | 44.0 | 41.4 | 41.1 | 40.5 |

**Table 1**. ML-MAP and MMI-MAP adaptation of ML and MMI systems from Switchboard: WERs on VM

ment from using an initially MMI-trained HMM set is retained if adaptation is with MMI-MAP but is partly lost with ML-MAP, especially with increasing adaptation data.

These results are also given in Table 1. There is 7.5% relative improvement from ML to MMI on the Switchboard-trained HMM set; the difference between ML-MAP-adapted ML and MMI-MAP-adapted MMI with 30h adaptation data is 8.0% relative. So the total improvement from discriminative training is 8.0%. Starting from the MMI-trained model, the improvement from using discriminative adaptation rather than ML adaptation is 4.6% relative.

| Training, Adaptation: | Hours adaptation data 0 | 1 | 4 | 15 | 20 | 30 |
|---|---|---|---|---|---|---|
| MPE,ML-MAP | 46.0 | 44.8 | 43.6 | 42.9 | 42.5 | 42.6 |
| MPE,MMI-MAP | 46.0 | 44.9 | 43.3 | 41.1 | 40.6 | 40.2 |

**Table 2**. ML-MAP and MMI-MAP adaptation of MPE-trained system from Switchboard: WERs on VM

Table 2 gives results from both MMI-MAP and ML-MAP adaptation starting from an MPE-trained system. The initial MPE-trained system is better than the MMI-trained system by 1.0% absolute, but after 30h of adaptation data this advantage over MMI is reduced to 0.3% absolute with MMI-MAP, or 0.0% with ML-MAP. The best results are obtained by using MMI-MAP to adapt an MPE-trained system (40.2%). Other experiments investigated the MPE-MAP adaptation of an MPE-trained system using a similar approach to MMI-MAP, but this did not robustly give improved results. There may not be enough adaptation data for MPE to give better results than MMI.

Experiments are reported in [2] in which Switchboard and Voicemail data are used together to train HMMs using MMI and ML, weighting the 30h of Voicemail data by varying amounts. With MMI training, the optimal weight is 2; WER is 41.6% with combined data. Using this as a baseline, MMI followed by MMI-MAP gives a 2.6% relative improvement to 40.5%. With ML training the optimal weight is 10 and the WER is 44.5%; relative to this baseline, ML followed by ML-MAP gives a 1.1% relative improvement to 44.0. So for adaptation, MAP training is better than training with combined weighted data.

## 5. CONCLUSIONS

A method has been described for MAP adaptation of HMM sets using the MMI criterion. This has been shown to be effective in maintaining the relative improvement of MMIE over MLE in the context of task adaptation. Furthermore the technique could also be applied to models trained which used the MPE criterion for both initial models and/or discriminative adaptation. The theory behind this form of MAP also provides a justification for the technique of I-smoothing [6] as a method of discriminative training with prior information.

While this paper has evaluated MMI-MAP in the context of task adaptation it could also be applied to speaker adaptation with large amounts of enrolment data or to the creation of discriminatively trained gender-dependent models using adaption techniques.

## 6. REFERENCES


[1] R. Cordoba, P.C. Woodland & M.J.F. Gales (2002). "Improved Cross-task Recognition Using MMIE Training," *ICASSP'02*, Orlando, Florida.

[2] M.J.F. Gales, Y. Dong, D. Povey & P.C. Woodland (2003). "Porting: Switchboard to the Voicemail Task", Submitted to *ICASSP'03*, Hong Kong.

[3] Y. Gao, B. Ramabhadran, M. Picheny (2000). "New Adaptation Techniques for Large Vocabulary Continuous Speech Recognition," Proc. ICSA ITRW ASR2000, Paris.

[4] J. Gauvain, J & C. Lee (1994). "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains." *IEEE Transactions on Acoustics, Speech and Signal Processing 1994, Vol.2, no. 2*, pp. 291-299

[5] M. Padmanabhan, B. Ramabhadran, E. Eide, G. Ramaswamy, L.R. Bahl, P.S. Gopalakrishnan & S. Roukos (1997). "Transcription of new speaking styles– Voicemail." *Proc. DARPA Hub4 Workshop.*

[6] D. Povey & P.C. Woodland (2002). "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," *ICASSP'02*, Orlando, Florida.

[7] L.F. Uebel & P.C. Woodland (2001). Discriminative Linear Transforms for Speaker Adaptation. *Proc. ISCA ITR-Workshop on Adaptation Methods in Speech Recognition*, Sophia-Antipolis.

[8] P.C. Woodland & D. Povey (2002). "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition," *Computer Speech & Language* v. 16, no 1, pp. 25-48, Jan 2002.