AFFINE INVARIANT SPARSE MAXIMUM A POSTERIORI ADAPTATION

Peder A. Olsen, Jing Huang, Steven J. Rennie, Vaibhava Goel

Department of Speech and Language Algorithms,
T. J. Watson Research Center, IBM

ABSTRACT

Modern speech applications utilize acoustic models with billions of parameters, and serve millions of users. Storing an acoustic model for each user is costly. We show through the use of sparse regularization, that it is possible to obtain competitive adaptation performance by changing only a small fraction of the parameters of an acoustic model. This allows for the compression of speaker-dependent models: a capability that has important implications for systems with millions of users. We achieve a performance comparable to the best Maximum A Posteriori (MAP) adaptation models while only adapting 5% of the acoustic model parameters. Thus it is possible to compress the speaker dependent acoustic models by close to a factor of 20. The proposed sparse adaptation criterion improves three aspects of previous work: It combines ℓ_0 and ℓ_1 penalties, have different adaptation rates for mean and variance parameters and is invariant to affine transformations.

Index Terms— Bayesian prior, elastic net, non-smooth optimization.

1. INTRODUCTION

Sparse representations are a powerful emerging model for complex signals such as speech, [1, 2, 3]. The idea of using a sparse regularizer with Maximum A Posteriori (MAP) adaptation is not new, [4]. Our previous paper on sparse adaptation used the same adaptation rate for the mean and variance parameters. Unfortunately, MAP adaptation for tasks with amounts of data in the 10 minutes to 10 hours range attain the best performance by aggressively adapting mean parameters while not adapting the variances at all. This paper addresses the problem of adapting means and variances at different rates, and also combines the ℓ_1 and counting norm penalties into a single combined sparse regularizer. Finally, the ℓ_1 penalty is modified so that the sparsity structure does not change if both the underlying training and adaptation data are simultaneously scaled and shifted (an affine transform).

2. A BRIEF REVIEW OF MAP

We shall use $\Xi = \{\boldsymbol{\mu}_g, \mathbf{v}_g, \omega_g\}_{g=1}^G$ to denote the parameters of our acoustic model, where $\boldsymbol{\xi}_g = (\boldsymbol{\mu}_g, \mathbf{v}_g)$ are the mean and variance parameters associated with gaussian component g. Maximum A Posteriori adaptation, [5], uses a Bayesian prior, $P(\Xi) = \prod_g P(\boldsymbol{\xi}_g)$ to smooth the maximum likelihood (ML) estimate. The Bayesian likelihood can then be maximized using the following per component auxilliary log likelihood

$$L(\Xi) = \sum_t \gamma_g(\mathbf{x}_t) \log\left(\omega_g \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right) + \log P(\boldsymbol{\xi}_g), \quad (1)$$

where we have used $\gamma_g(\mathbf{x}_t) = \mathrm{P}(g|\Xi^{\mathrm{old}}, \mathbf{x}_t)$ for the gaussian posterior at time t. It has become quite common in speech recognition to assume diagonal covariance models $\boldsymbol{\Sigma}_g = \mathrm{diag}(\mathbf{v}_\mathrm{g})$. If the Bayesian prior also decouples across dimensions then the objective function can be written $L(\Xi) = \sum_{k=1}^d L(\{\omega_g, \mu_{kg}, v_{kg}\}_{g=1}^G)$, and the parameters in each dimension can be estimated independently. We therefore make the simplifying assumption that $\boldsymbol{\mu}_g = \mu_g$ and $\mathbf{v}_g = v_g$ are one dimensional scalars in the rest of the paper.
The Bayesian prior $P(\boldsymbol{\xi}_g)$ in [5] was chosen to be a conjugate prior given by the normal-Wishart distribution:

$$\begin{aligned} P(\boldsymbol{\xi}_g) &\stackrel{\mathrm{def}}{=} R^{\mathrm{MAP}}(\boldsymbol{\xi}_g|\boldsymbol{\xi}_g^{\mathrm{old}}) \\ &= P(\mu_g, v_g|\mu_g^{\mathrm{old}}, v_g^{\mathrm{old}}, \tau_\mu, \tau_v)\mathrm{P}(\omega_g|\omega_g^{\mathrm{old}}, \tau_\omega) \\ &\propto \left(\frac{e^{-\frac{(\mu-\mu^{\mathrm{old}})^2}{2v}}}{\sqrt{2\pi v}}\right)^{\tau_\mu} \left(\frac{v^{-\alpha}e^{-\frac{v^{\mathrm{old}}}{v}}}{(v^{\mathrm{old}})^{-\alpha-1}\Gamma(\alpha+1)}\right)^{\tau_v} \left(\omega_g^{\omega_g^{\mathrm{old}}}\right)^{\tau_\omega}. \end{aligned}$$

This prior makes the parameter estimation straightforward. In the case of no Bayesian prior ($\tau_\mu = \tau_v = \tau_\omega = 0$) the ML solution is

$$\omega_g^{\mathrm{ML}} = \frac{1}{T}\sum_{t=1}^T \gamma_g(x_t), \qquad \mu_g^{\mathrm{ML}} = \frac{1}{T_g}\sum_{t=1}^T \gamma_g(x_t)x_t, \quad (2)$$

$$v_g^{\mathrm{ML}} = \frac{1}{T_g}\sum_{t=1}^T \gamma_g(x_t)x_t^2 - (\mu_g^{\mathrm{ML}})^2 \quad (3)$$

where $T_g = T\omega_g^{\mathrm{ML}} = \sum_{t=1}^T \gamma_g(x_t)$. In terms of the maximum likelihood solution the objective functions can be written

$$\begin{aligned} L(\boldsymbol{\xi}_g) &= -\frac{T_g}{2}\left(\frac{(\mu_g - \mu_g^{\mathrm{ML}})^2 + v_g^{\mathrm{ML}}}{v_g} + \log(v_g)\right) \quad (4) \\ &\quad +T_g\log(\omega_g) + \log P(\boldsymbol{\xi}_g) \\ &= -(T_g+\tau_\mu)\frac{(\mu_g - \mu_g^{\mathrm{MAP}})^2}{2v_g} \quad (5) \\ &\quad -(T_g+\tau_\mu+\alpha\tau_v)\left(\frac{v_g^{\mathrm{MAP}}}{2v_g} + \frac{1}{2}\log(v_g)\right) \\ &\quad +(T\omega_g^{\mathrm{ML}} + \tau_\omega\omega_g^{\mathrm{old}})\log\omega_g, \end{aligned}$$

where

$$\omega_g^{\mathrm{MAP}} = \frac{T\omega_g^{\mathrm{ML}} + \tau_\omega\omega_g^{\mathrm{old}}}{T+\tau_\omega} \quad (6)$$

$$\mu_g^{\mathrm{MAP}} = \frac{T_g\mu_g^{\mathrm{ML}} + \tau_\mu\mu_g^{\mathrm{old}}}{T_g+\tau_\mu} \quad (7)$$

$$v_g^{\mathrm{MAP}} = \frac{\frac{T_g\tau_\mu}{T_g+\tau_\mu}(\mu_g^{\mathrm{ML}} - \mu_g^{\mathrm{old}})^2 + T_g v_g^{\mathrm{ML}} + \tau_v v_g^{\mathrm{old}}}{T_g+\tau_\mu+\alpha\tau_v}. \quad (8)$$

It is straightforward to verify that μ^{MAP} and v^{MAP} are values that maximizes $L(\boldsymbol{\xi}_g)$.
In our previous paper, [4], we chose the hyper-parameters to be $\tau_\mu = \tau_v = \tau_\omega = \tau, \alpha = 0$. This makes the Bayesian log-prior particularly simple and MAP adaptation in this case corresponds to I-smoothing as described in [6]. Unfortunately, for the best performance we had to sacrifice simplicity. We will see in the experiment section that the best results for MAP adaptation on our data was obtained for $\tau_\mu = \tau_\omega = \tau, \tau_v = \infty, \alpha = 1$.

3. SPARSITY PROMOTING BAYESIAN PRIORS

In order to compactly store the speaker specific acoustic model we can store the differences $\boldsymbol{\xi}_g - \boldsymbol{\xi}_g^{\mathrm{old}}$ instead of the models themselves ($\boldsymbol{\xi}_g$). Since the parameters change very little we can save storage by storing only the significant differences. By using a Bayesian prior with sparse regularisation, only the significant parameters will be allowed to have non-zero values. We shall consider the following such Bayesian prior:

$$P(\boldsymbol{\xi}_g) \propto R^{\mathrm{MAP}}(\boldsymbol{\xi}_g|\boldsymbol{\xi}_g^{old}) R^{\mathrm{sparse}}(\boldsymbol{\xi}_g|\boldsymbol{\xi}_g^{\mathrm{old}})$$

where

$$\begin{aligned}\log R^{\mathrm{sparse}}(\boldsymbol{\xi}_g|\boldsymbol{\xi}_g^{\mathrm{old}}) &= -\lambda_{1\mu}|\mu_g - \mu_g^{\mathrm{old}}| - \lambda_{1v}|v_g - v_g^{\mathrm{old}}| \\ &\quad -\lambda_{0\mu}\|\mu_g - \mu_g^{\mathrm{old}}\|_0 - \lambda_{0v}\|v_g - v_g^{\mathrm{old}}\|_0,\end{aligned} \tag{9}$$

and $\|x\|_0$ is defined to be 0 if $x = 0$ and 1 otherwise. The first part is a weighted ℓ_1 penalty and the second part is a weighted count of the parameter changes. This sparse regularizer is not differentiable or continuous, but it is piecewise continuous and differentiable. For example it is continuous in μ, v on the quadrant $\mu > \mu^{\mathrm{old}}$, $v > v^{\mathrm{old}}$, and likewise continuous in μ on the line segment $\mu < \mu^{\mathrm{old}}$, $v = v^{\mathrm{old}}$. Therefore we can find all local maxima by considering separately the 9 different pieces for which the function is continuous. We will show how to find the global maximum in the next section.

3.1. Affine Invariance

If we change the adaptation data by a linear transform $x_t \to ax_t + b$ then the corresponding maximum likelihood estimates changes in a predictable manner: $\mu_g^{\mathrm{ML}} \to a\mu_g^{\mathrm{ML}} + b$, $v_g^{\mathrm{ML}} \to a^2 v_g^{\mathrm{ML}}$. Likewise if the training data undergoes the same linear transform then the base acoustic model will be changed similarly: $\mu_g^{\mathrm{old}} \to a\mu_g^{\mathrm{old}} + b$, $v_g^{\mathrm{old}} \to a^2 v_g^{\mathrm{old}}$. We will say that an adaptation model is *invariant to affine transforms* if the adapted model undergoes the same transform. This is a property that the MAP adaptation estimate satisfies: $\mu_g^{\mathrm{MAP}} \to a\mu_g^{\mathrm{MAP}} + b$, $v_g^{\mathrm{MAP}} \to a^2 v_g^{\mathrm{MAP}}$. We would like the same to be true for the sparse estimate. For that to be the case we need the sparse regularizer $R^{\mathrm{sparse}}(\boldsymbol{\xi}_g|\boldsymbol{\xi}_g^{\mathrm{old}})$ to be invariant to affine transformation. The proposed sparse regularizer (9) is not invariant to scaling, but the following modified regularizer is now scale-invariant and therefore also invariant to affine transformations:

$$\begin{aligned}\log R^{\mathrm{sparse}}(\boldsymbol{\xi}_g|\boldsymbol{\xi}_g^{\mathrm{old}}) &= -\frac{\lambda_{1\mu}}{\sqrt{v_g^{\mathrm{old}}}}|\mu_g - \mu_g^{\mathrm{old}}| - \frac{\lambda_{1v}}{v_g^{\mathrm{old}}}|v_g - v_g^{\mathrm{old}}| \\ &\quad -\lambda_{0\mu}\|\mu_g - \mu_g^{\mathrm{old}}\|_0 - \lambda_{0v}\|v_g - v_g^{\mathrm{old}}\|_0,\end{aligned}$$

If we use the last sparse regularizer together with the general MAP regularizer we will have corrected the three shortcomings of our previous paper. The described regularizer gives affine invariance, a combined counting norm and ℓ_1 norm regularizer and different adaptation rates for the mean and variance parameters.

4. ANALYTIC OPTIMIZATION

Despite the fact that neither the function nor its derivative is continuous everywhere, we can still find the global optimum analytically. This remarkable fact ensures that there is very little added computational cost to use a sparse regularizer. We show how to find the analytic solution in this section. Instead of maximizing the Bayesian likelihood, $L(\boldsymbol{\xi}_g)$, we minimize the negative log likelihood, $F(\boldsymbol{\xi}_g) = -L(\boldsymbol{\xi}_g)$. Omitting terms related to ω_g, the objective function $F(\mu_g, v_g)$ can be written:

$$\begin{aligned}F(\boldsymbol{\xi}_g) &= F(\boldsymbol{\xi}_g; \boldsymbol{\xi}_g^{\mathrm{old}}, \boldsymbol{\xi}_g^{\mathrm{ML}}, \tau_\mu, \tau_v, \alpha, \lambda_{0\mu}, \lambda_{0v}, \lambda_{1\mu}, \lambda_{1v}) \\ &= (T_g + \tau_\mu)\frac{(\mu_g - \mu_g^{\mathrm{MAP}})^2}{2v_g} \\ &\quad +(T_g + \tau_\mu + \alpha\tau_v)\left(\frac{v_g^{\mathrm{MAP}}}{2v_g} + \frac{1}{2}\log(v_g)\right) \\ &\quad +\frac{\lambda_{1\mu}}{\sqrt{v_g^{\mathrm{old}}}}|\mu_g - \mu_g^{\mathrm{old}}| + \frac{\lambda_{1v}}{v_g^{\mathrm{old}}}|v_g - v_g^{\mathrm{old}}| \\ &\quad +\lambda_{0\mu}\|\mu_g - \mu_g^{\mathrm{old}}\|_0 + \lambda_{0v}\|v_g - v_g^{\mathrm{old}}\|_0,\end{aligned} \tag{10}$$

Let's consider a local minimum of F occuring in the interior of the lower quadrant defined by $\mu_g < \mu_g^{\mathrm{old}}$, $v_g < v_g^{\mathrm{old}}$. Defining $T_\mu = (T_g + \tau_\mu)/2$, $T_v = (T_g + \tau_\mu + \alpha\tau_v)/2$, $\hat{\lambda}_{1\mu} = \lambda_{1\mu}/\sqrt{v_g^{\mathrm{old}}}$ and $\hat{\lambda}_{1v} = \lambda_{1v}/v_g^{\mathrm{old}}$ the function on this quadrant equals:

$$\begin{aligned}F(\xi_g) &= T_\mu\frac{(\mu_g - \mu_g^{\mathrm{MAP}})^2}{v_g} + T_v\left(\frac{v_g^{\mathrm{MAP}}}{v_g} + \log(v_g)\right) \\ &\quad -\hat{\lambda}_{1\mu}(\mu_g - \mu_g^{\mathrm{old}}) - \hat{\lambda}_{1v}(v_g - v_g^{\mathrm{old}}) + \lambda_{0\mu} + \lambda_{0v}.\end{aligned} \tag{11}$$

The corresponding partial derivatives are

$$\begin{aligned}\frac{\partial F}{\partial \mu_g} &= 2T_\mu\frac{(\mu_g - \mu_g^{\mathrm{MAP}})}{v_g} - \hat{\lambda}_{1\mu} \\ \frac{\partial F}{\partial v_g} &= -T_\mu\frac{(\mu_g - \mu_g^{\mathrm{MAP}})^2}{v_g^2} + T_v\left(\frac{-v_g^{\mathrm{MAP}}}{v_g^2} + \frac{1}{v_g}\right) - \hat{\lambda}_{1v}.\end{aligned}$$

The local minimum can be found by solving the critical equations. The first derivative is zero when $(\mu_g - \mu_g^{\mathrm{MAP}})/v_g = \hat{\lambda}_{1\mu}/(2T_\mu)$. If we substitute this into the expression for the second derivative, then the second derivative is zero when

$$0 = -T_\mu\left(\frac{\hat{\lambda}_{1\mu}}{2T_\mu}\right)^2 + T_v\left(\frac{-v_g^{\mathrm{MAP}}}{v_g^2} + \frac{1}{v_g}\right) - \hat{\lambda}_{1v},$$

or equivalently

$$0 = v_g^{\mathrm{MAP}} - v_g + v_g^2\left(\frac{(\hat{\lambda}_{1\mu})^2}{4T_\mu T_v} + \frac{\hat{\lambda}_{1v}}{T_v}\right) \tag{12}$$

If we define

$$\beta = \left(\frac{(\hat{\lambda}_{1\mu})^2}{4T_\mu T_v} + \frac{\hat{\lambda}_{1v}}{T_v}\right)$$

the two solutions to (12) are:

$$v_g = \frac{1 \pm \sqrt{1 - 4\beta v_g^{\mathrm{MAP}}}}{2\beta}. \tag{13}$$

It can be seen that the first of smallest of these two roots corresponds to a local minimum since the derivative with respect to v_g goes to

$-\infty$ as $v_g \to 0$. The second root corresponds to a local maximum. We conclude that

$$v_g = \frac{1-\sqrt{1-4\beta v_g^{\mathrm{MAP}}}}{2\beta}$$

$$\mu_g = \mu_g^{\mathrm{MAP}} + \frac{\hat{\lambda}_{1\mu} v_g}{2T_\mu}$$

is a valid solution if $4\beta v_g^{\mathrm{MAP}} \leq 1$, $v_g < v_g^{\mathrm{old}}$ and $\mu_g < \mu_g^{\mathrm{old}}$. The solution for the other regions are no more difficult. The complete solution is given in Algorithm 1.

input : Statistics $\mathbf{s}$, T_g, model μ_g^{old}, v_g^{old} and penalty parameters $\tau_\mu, \tau_v, \alpha, \lambda_{0\mu}, \lambda_{0v}, \lambda_{1\mu}, \lambda_{1v}$
output: Global minimum $\mu_{\min}, v_{\min}$

Compute: $\mu_g^{\mathrm{ML}} = s_1$, $v_g^{\mathrm{ML}} = s_2 - s_1^2$
$\mu_g^{\mathrm{MAP}} = \frac{T_g \mu_g^{\mathrm{ML}} + \tau_\mu \mu_g^{\mathrm{old}}}{T_g + \tau_\mu}$
$v_g^{\mathrm{MAP}} = \frac{\frac{T_g \tau_\mu}{T_g + \tau_\mu}(\mu_g^{\mathrm{ML}} - \mu_g^{\mathrm{old}})^2 + T_g v_g^{\mathrm{ML}} + \tau_v v_g^{\mathrm{old}}}{T_g + \tau_\mu + \alpha\tau_v}$
$T_1 = (T_g + \tau_\mu)/2$, $T_2 = (T_g + \tau_\mu + \alpha\tau_v)/2$
Update: $\lambda_{1\mu} \leftarrow \frac{\lambda_{1\mu}}{\sqrt{v_g^{\mathrm{old}}}}$, $\lambda_{1v} \leftarrow \frac{\lambda_{1v}}{v_g^{\mathrm{old}}}$
$F_{\mathrm{test}} = F(\mu_g^{\mathrm{old}}, v_g^{\mathrm{old}})$
Initialize: $(F_{\min}, \mu_{\min}, v_{\min}) = (F_{\mathrm{test}}, \mu_g^{\mathrm{old}}, v_g^{\mathrm{old}})$
for $\delta \in \{-1, 1\}$ **do**
  $\mu_g = \mu_g^{\mathrm{MAP}} - \frac{\delta \lambda_{1\mu} v_g^{\mathrm{old}}}{2T_1}$
  if $\delta(\mu_g - \mu_g^{\mathrm{old}}) > 0$ **then**
    $F_{\mathrm{test}} = F(\mu_g, v_g^{\mathrm{old}})$
    if $F_{\mathrm{test}} < F_{\min}$ **then**
      $(F_{\min}, \mu_{\min}, v_{\min}) = (F_{\mathrm{test}}, \mu_g, v_g^{\mathrm{old}})$
  end
end
for $\epsilon \in \{-1, 1\}$ **do**
  $v_g^* = \frac{T_1}{T_2}(\mu_g^{\mathrm{old}} - \mu_g^{\mathrm{MAP}})^2 + v_g^{\mathrm{MAP}}$
  if $1 + 4\epsilon v_g^* \lambda_{1v}/T_2 \geq 0$ **then**
    $v_g = \frac{1-\sqrt{1+4\epsilon v_g^* \lambda_{1v}/T_2}}{-2\epsilon\lambda_{1v}/T_2}$
    if $\epsilon(v_g - v_g^{\mathrm{old}}) > 0$ **then**
      $F_{\mathrm{test}} = F(\mu_g^{\mathrm{old}}, v_g)$
      if $F_{\mathrm{test}} < F_{\min}$ **then**
        $(F_{\min}, \mu_{\min}, v_{\min}) = (F_{\mathrm{test}}, \mu_g^{\mathrm{old}}, v_g)$
    end
  end
end
for $\epsilon \in \{-1, 1\}$ **do**
  $\beta = \frac{\lambda_{1\mu}^2 - 4\epsilon T_1 \lambda_{1v}}{4T_1 T_2}$
  if $1 - 4\beta v_g^{\mathrm{MAP}} \geq 0$ **then**
    for $\delta \in \{-1, 1\}$ **do**
      $v_g = \begin{cases} v_g^{\mathrm{MAP}} & \text{if } \beta = 0 \\ \frac{1-\sqrt{1-4\beta v_g^{\mathrm{MAP}}}}{2\beta} & \text{if } \beta \neq 0 \end{cases}$
      $\mu_g = \mu_g^{\mathrm{MAP}} - \delta \frac{\lambda_{1\mu}}{2T_1} v_g$
      if $\delta(\mu_g - \mu_g^{\mathrm{old}}) > 0$ **and** $\epsilon(v_g - v_g^{\mathrm{old}}) > 0$ **then**
        $F_{\mathrm{test}} = F(\mu_g, v_g)$
        if $F_{\mathrm{test}} < F_{\min}$ **then**
          $(F_{\min}, \mu_{\min}, v_{\min}) = (F_{\mathrm{test}}, \mu_g, v_g)$
      end
    end
  end
end

Algorithm 1: The global minimum of (10).

5. EXPERIMENTS

5.1. Task Description

We used an internal US English speech recognition task for all experiments. The training set consists of 2000 hours of recordings. The test and adaptation sets were collected from the same set of 26 speakers. The enrollment data used for adaptation consists of 2.8 ± 1.5 hours of data per speaker with known transcripts. The test data has 7-20 minutes per speaker (52K words in total). Acoustic features were constructed from 12 dimensional Mel-frequency Cepstra coefficients and their first, second and third derivative, followed by a Linear Discriminant Analysis (LDA) projection.
The acoustic model had 5000 HMM states and 200,000 gaussian components. A Constrained Maximum Likelihood Linear Regression (CMLLR) transform, [7], was learned for each speaker in the training database. In the transformed feature space, we then trained a "canonical acoustic model" using speaker adaptive training (SAT), [8]. At test time we only used the canonical acoustic model. The canonical acoustic model was trained using feature space minimum phone error rate (fMPE) and discriminative Minimum Phone Error (MPE), as described in [9, 10].

5.2. Baseline MAP Results

Table 1 shows the baseline word error rates (WER) with and without MAP adaptation. The baseline WER for Ξ^{old} was 13.2%, which with a typical MAP adaptation setting ($\tau_\mu = \tau_v = 100$, $\alpha = 0$) gave a 17% reduction of the WER to 10.9%. The best MAP adaptation results was 10.5%, and was attained by not adapting any of the variance parameters ($\tau_\mu = 100$, $\tau_v = \infty$, $\alpha = 1$). As we did not want to compromise on accuracy, this model became the baseline MAP model for all our experiments with sparsity. Table 1 also shows results when we use maximum likelihood estimation to create the new model. Since there are many states for which there is only one or two data points, the variance estimate becomes particularly unreliable (for states with no data, we leave the gaussians unchanged). Maximum likelihood estimation for all the parameters almost doubles the WER, while only applying maximum likelihood estimation to means gives a WER 11.2%.
In Table 2 we can see results for various settings of τ_μ, τ_v and α. As seen in the table, the best WER is reached when $\tau_v = \infty$. In other words we were not able to extract any useful information from the variance parameters. We did not investigate whether this was due to the small amount of data or the fact that the base acoustic model was trained using a discriminative objective function.

5.3. Sparse MAP Adaptation Results

In Table 3 we give the word error rate and percentage of parameters that do not change (we call this the sparsity). Without any sparsity

System	WER
SAT+FMPE+CMLLR (no adaptation)	13.2%
ML (μ_g and v_g)	23.8%
ML (μ_g only)	11.2%
MAP $\tau_\mu = \tau_v = 100$, $\alpha = 0$	10.9%
MAP, $\tau_\mu = 100$**,** $\tau_v = \infty$**,** $\alpha = 1$	**10.5%**

Table 1. Word error rates for baseline systems

τ_μ	τ_v	α	Sparsity	WER
50	50	0	2%	11.3%
100	100	0	2%	10.9%
500	500	0	2%	10.9%
1000	1000	0	2%	11.2%
100	100	1	2%	10.9%
100	1000	0.9	2%	10.7%
100	10^6	1	2%	10.5%
100	∞	1	52%	10.5%

Table 2. Word error rates for MAP systems for various values of τ_μ, τ_v and α.

promoting penalty MAP results in 52% sparsity – all the variances are left unchanged and about 2% of the gaussians corresponded to states with no data. If we try to optimize the word error rate with respect to all of the parameters τ_μ, τ_v, α, $\lambda_{0\mu}$, λ_{0v}, $\lambda_{1\mu}$ and λ_{1v} then the best word error rate we can achieve is 10.4%. This system used both the MAP, ℓ_1 and ℓ_0 parts of the penalty: $\tau_\mu = 50$, $\lambda_{0\mu} = 0.1$, $\lambda_{1\mu} = 0.3$. We don't believe, however, that this is a significant improvement over the MAP baseline. To verify that the sparse penalty with affine invariance is better than (9) we repeated this experiment with (9). To match the 84% sparsity we found the corresponding $\lambda_{1\mu} = 0.02$ gave a WER of 10.5%. We found this to be consistently the case – that the affine invariant penalty was marginally better than the sparse penalty (9).

The highest sparsity we could obtain without degrading the WER was 92%. This system used $\tau_\mu = 100$, $\lambda_{0\mu} = 0.3$, $\lambda_{1\mu} = 1$. The table also lists the best system we could get at a sparsity level of 95% and 98%. For 95% sparsity the optimal system could be obtained without activating the ℓ_1 penalty. This system had a WER of 10.6% – a small degradation. For the more aggressive 98% sparsity level the WER was 10.8%.

6. DISCUSSION

We have shown through the use of sparse regularization, that it is possible to obtain competitive MAP adaptation performance by changing only a small fraction of the parameters of an acoustic model. This allows for the compression of speaker-dependent models: a capability that has important implications for systems with millions of users. At a 95% sparsity level the speaker dependent acoustic models could be compressed by a maximum factor of 20. If the actual locations of the changed parameters can be stored efficiently we can hope to get close to the factor of 20. Since all the variances are unchanged and a large fraction of the gaussians remain unchanged, it is only a small fraction of the gaussians for which the change locations need to be encoded. The entropy of the binary mask for the remaining parameter change locations suggests

τ_μ	$\lambda_{0\mu}$	λ_{0v}	$\lambda_{1\mu}$	λ_{1v}	Sparsity	WER
100	0	0	0	∞	52%	10.5%
100	0	0	0.3	∞	73%	10.5%
100	0	0	1	∞	83%	10.6%
50	0	0	0.3	∞	73%	10.5%
15	0	0	0.3	∞	73%	10.6%
50	0.1	0	0.3	∞	84%	10.4%
25	0.3	0	1	∞	92%	10.5%
100	1.0	∞	0	0	95%	10.6%
50	2.0	0	0.3	∞	98%	10.8%

Table 3. Word error rates for sparse MAP systems for various values of τ_μ, $\lambda_{0\mu}$, λ_{0v}, $\lambda_{1\mu}$ and λ_{1v}. We used $\alpha = 1$ and $\tau_v = 0$ in all these experiments.

that a Huffman coding, [11], would obtain a compression factor close to 20. Future work should compare and potentially combine sparse methods with parameter tying methods such as CMLLR with multiple transforms.

7. REFERENCES

[1] M. Zibulevsky and M. Elad, "L1-L2 optimization in signal and image processing," *Signal Processing Magazine, IEEE*, vol. 27, no. 3, pp. 76–88, 2010.

[2] G. Sivaram, S.K. Nemala, M. Elhilali, T.D. Tran, and H. Hermansky, "Sparse coding for speech recognition," in *ICASSP*. IEEE, 2010, pp. 4346–4349.

[3] T.N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *ICASSP*. IEEE, 2010, pp. 4370–4373.

[4] P. A. Olsen, J. Huang, S. J. Rennie, and V. Goel, "Sparse maximum a posteriori adaptation," in *ASRU*, 2011.

[5] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[6] D. Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D. thesis, Cambridge University, 2003.

[7] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[8] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *ICSLP*, 1996, vol. 2, pp. 1137–1140.

[9] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *ICASSP*, 2005, vol. 1, pp. 961–964.

[10] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *ICASSP*, 2002, vol. I, pp. 105–108.

[11] D.A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.