Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition

Lu, L., Ghoshal, A., & Renals, S. (2012). Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. (pp. 4877-4880). IEEE. doi: 10.1109/ICASSP.2012.6289012

MAXIMUM A POSTERIORI ADAPTATION OF SUBSPACE GAUSSIAN MIXTURE MODELS FOR CROSS-LINGUAL SPEECH RECOGNITION

Liang Lu[1]*, Arnab Ghoshal*[1,2]*, and Steve Renals*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, UK
[2]Spoken Language Systems, Saarland University, Germany
{liang.lu, a.ghoshal, s.renals}@ed.ac.uk

ABSTRACT

This paper concerns cross-lingual acoustic modeling in the case when there are limited target language resources. We build on an approach in which a subspace Gaussian mixture model (SGMM) is adapted to the target language by reusing the globally shared parameters estimated from out-of-language training data. In current cross-lingual systems, these parameters are fixed when training the target system, which can give rise to a mismatch between the source and target systems. We investigate a maximum a posteriori (MAP) adaptation approach to alleviate the potential mismatch. In particular, we focus on the adaptation of phonetic subspace parameters using a matrix variate Gaussian prior distribution. Experiments on the GlobalPhone corpus using the MAP adaptation approach results in word error rate reductions, compared with the cross-lingual baseline systems and systems updated using maximum likelihood, for training conditions with 1 hour and 5 hours of target language data.

Index Terms— Subspace Gaussian Mixture Model, Maximum a Posteriori Adaptation, Cross-lingual Speech Recognition

1. INTRODUCTION

In the subspace Gaussian mixture model (SGMM) [1], the model parameters are derived from a set of state dependent parameters, and from a set of globally shared parameters which capture the phonetic and speaker variation. This is in contrast to conventional HMM/GMM based speech recognition systems in which the state model parameters are estimated directly. Decoupling the globally shared and state-specific parameters results in a decrease in the total number of free parameters in the model. Additionally, it is possible to estimate the global parameters using out-of-domain or out-of-language data when there is limited labeled acoustic data for the target domain or language.

This idea has been explored in the application of SGMMs to multilingual speech recognition [2], where the globally shared parameters were estimated by tying across multiple languages to improve estimation accuracy. It has also been used in cross-lingual settings [2, 3], where the global parameters were reused by the target language system, with only state dependent parameters being re-estimated. Experiments have shown that significant performance improvements could be achieved when training data for the target language is very limited, since the number of parameters to be estimated is much smaller [3].

The research leading to these results was supported by European Community's Seventh Framework Programme under grant agreement number 213850 (SCALE), and by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

However, sharing the global parameter set across multiple languages can introduce a mismatch with the target language system, owing to differences in phonetic characteristics, corpus recording conditions, and speaking styles. Since the amount of training data may not be sufficient to allow the global parameters to be updated using maximum likelihood (ML), in this paper we employ maximum a posteriori (MAP) adaptation. In particular, we train the target language system using MAP adaptation of the phonetic subspace parameters with a matrix variate Gaussian prior distribution based on the phonetic subspace parameters estimated in the multilingual system. In experiments with a cross-lingual framework [3], we have observed that the MAP adaptation approach results in a considerable reduction in word error rate (WER).

2. SGMM ACOUSTIC MODEL

In an SGMM [1], the HMM state is modelled as:

$$P(\mathbf{o}_t|j,s) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi}\, \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jmi}^{(s)}, \boldsymbol{\Sigma}_i), \quad (1)$$

$$\boldsymbol{\mu}_{jmi}^{(s)} = \mathbf{M}_i \mathbf{v}_{jm} + \mathbf{N}_i \mathbf{v}^{(s)}, \quad (2)$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}, \quad (3)$$

where $\mathbf{o}_t \in \mathbb{R}^D$ denotes the t-th D-dimensional acoustic frame, j is the HMM state index, m is a sub-state [1], I is the number of Gaussians, and $\boldsymbol{\Sigma}_i$ is the i-th (globally shared) covariance matrix. $\mathbf{v}_{jm} \in \mathbb{R}^S$ is referred to as the sub-state vector, and S denotes the subspace dimension. The matrices $\mathbf{M}_i$ and the vectors $\mathbf{w}_i$ span the model subspaces for the Gaussian means and weights respectively, and are used to derive the GMM parameters given sub-state vectors (equations (2) and (3)). Similarly, $\mathbf{N}_i$ defines the speaker subspace for Gaussian means, and $\mathbf{v}^{(s)} \in \mathbb{R}^T$ is referred to as the speaker vector where T denotes the dimension of the speaker subspace.

3. MAP ESTIMATION OF THE PHONETIC SUBSPACE

In ML estimation of the phonetic subspace [1], the auxiliary function for $\mathbf{M}_i$ is given by:

$$\mathcal{Q}(\mathbf{M}_i) \propto tr(\mathbf{M}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{Y}_i) - \frac{1}{2} tr(\mathbf{M}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i \mathbf{Q}_i), \quad (4)$$

where

$$\begin{cases} \mathbf{Y}_i = \sum_{jmt} \tilde{\gamma}_{jmi}(t) \mathbf{o}_t \mathbf{v}_{jm}^T \\ \mathbf{Q}_i = \sum_{jm} \gamma_{jmi} \mathbf{v}_{jm} \mathbf{v}_{jm}^T \end{cases}, \quad (5)$$

$\tilde{\gamma}_{jmi}(t)$ denotes the Gaussian component posterior for acoustic frame $\mathbf{o}_t$, and $\gamma_{jmi} = \sum_t \tilde{\gamma}_{jmi}(t)$. If a prior term is introduced, then the auxiliary function becomes:

$$\tilde{\mathcal{Q}}(\mathbf{M}_i) = \mathcal{Q}(\mathbf{M}_i) + \tau \log P(\mathbf{M}_i), \tag{6}$$

where $P(\mathbf{M}_i)$ denotes the prior distribution of matrix $\mathbf{M}_i$, and τ is the smoothing parameter which balances the relative contributions of the likelihood and prior. Although any valid form of $P(\mathbf{M}_i)$ may be used, in practical MAP applications a conjugate prior distribution is often preferred for reasons of simplicity. In this paper, $P(\mathbf{M}_i)$ is set to be a Gaussian distribution which is conjugate to the auxiliary $\mathcal{Q}(\mathbf{M}_i)$.

3.1. Matrix Variate Gaussian Prior

The Gaussian distribution of random matrices is well understood [4]. A typical example of its application in speech recognition is maximum a posteriori linear regression (MAPLR) [5] for speaker adaptation, in which a matrix variate prior was used for the linear regression transformation matrix. The Gaussian distribution of a $D \times S$ matrix $\mathbf{M}$ is defined as:

$$\log P(\mathbf{M}) = -\frac{1}{2}\Big(DS\log(2\pi) + D\log|\mathbf{\Omega}_r| + S\log|\mathbf{\Omega}_c| + tr\big(\mathbf{\Omega}_r^{-1}(\mathbf{M}-\bar{\mathbf{M}})\mathbf{\Omega}_c^{-1}(\mathbf{M}-\bar{\mathbf{M}})^T\big)\Big), \tag{7}$$

where $\bar{\mathbf{M}}$ is a matrix containing the expectation of each element of $\mathbf{M}$, and $\mathbf{\Omega}_r$ and $\mathbf{\Omega}_c$ are $D \times D$ and $S \times S$ positive definite matrices representing the covariance between the rows and columns of $\mathbf{M}$, respectively. $|\cdot|$ and $tr(\cdot)$ denote the determinant and trace of a square matrix. This prior distribution is conjugate to auxiliary function (4). This matrix density Gaussian distribution may be written as:

$$Vec(\mathbf{M}) \sim \mathcal{N}(Vec(\bar{\mathbf{M}}), \mathbf{\Omega}_r \otimes \mathbf{\Omega}_c), \tag{8}$$

where $Vec(\cdot)$ is the vectorization operation which maps a $D \times S$ matrix into a $DS \times 1$ vector, and $\otimes$ denotes the Kronecker product of two matrices. In this formulation, only $\mathbf{\Omega}_r \otimes \mathbf{\Omega}_c$ is uniquely defined, and not the individual covariances $\mathbf{\Omega}_r$ and $\mathbf{\Omega}_c$, since for any $\alpha > 0$, $(\alpha\mathbf{\Omega}_r, \frac{1}{\alpha}\mathbf{\Omega}_c)$ would lead to the same distribution. However, this is not of concern in the current application to MAP adaptation.

3.2. Prior Distribution Estimation

For MAP estimation, the prior distribution $P(\mathbf{M}_i)$ for each $\mathbf{M}_i$, should be estimated first. This requires the estimation of the mean matrices $\bar{\mathbf{M}}_i$, and the row and column covariances $\mathbf{\Omega}_r$ and $\mathbf{\Omega}_c$. Given a set of samples generated by $P(\mathbf{M}_i)$, the ML estimation of the mean, and the row and column covariances, is described by Dutilleul [6]. In MAPLR such samples are derived from clusters in the speaker independent model based on a regression tree [5]. In the case of cross-lingual SGMMs, the MAP formulation is based on the assumption that the multilingual estimate of the global subspace parameters serves a good starting point, which has been empirically verified earlier [3]. Recall that in the current cross-lingual system, the subspace parameters are obtained from an initial multilingual system trained on the source languages, and fixed during training of the state-specific parameters on the target language data. For its MAP counterpart, we set these multilingual parameters to be the mean of the prior $P(\mathbf{M}_i)$ and update both the state-specific $\mathbf{v}_{jm}$ and the global $\mathbf{M}_i$. With a sufficiently large value of τ in (6), we can shrink the system back to the cross-lingual baseline, whereas $\tau = 0$ corresponds to the ML update.

The covariance matrices for each $P(\mathbf{M}_i)$ are global, estimated from the multilingual parameters by ML [6]. To be specific, suppose the set of multilingual phonetic subspace matrices is $\{\bar{\mathbf{M}}_i, i = 1, \ldots, I\}$. We first compute the global mean as $\bar{\mathbf{M}} = \frac{1}{I}\sum_{i=1}^{I}\bar{\mathbf{M}}_i$. The two covariance matrices, $\mathbf{\Omega}_r$ and $\mathbf{\Omega}_c$, are then estimated by computing the following two equations iteratively until convergence:

$$\begin{aligned}\mathbf{\Omega}_r &= \tfrac{1}{ID}\textstyle\sum_{i=1}^{I}\left(\bar{\mathbf{M}}_i - \bar{\mathbf{M}}\right)\mathbf{\Omega}_c^{-1}\left(\bar{\mathbf{M}}_i - \bar{\mathbf{M}}\right)^T \\ \mathbf{\Omega}_c &= \tfrac{1}{IS}\textstyle\sum_{i=1}^{I}\left(\bar{\mathbf{M}}_i - \bar{\mathbf{M}}\right)^T\mathbf{\Omega}_r^{-1}\left(\bar{\mathbf{M}}_i - \bar{\mathbf{M}}\right).\end{aligned} \tag{9}$$

$\mathbf{\Omega}_r, \mathbf{\Omega}_c$ can be initialised as identity matrices, and several iterations are found to be sufficient for convergence. Hence the prior $P(\mathbf{M}_i)$ is parameterized by $\bar{\mathbf{M}}_i$, $\mathbf{\Omega}_r$, and $\mathbf{\Omega}_c$.

Povey [7] has discussed using a global prior over all the subspace matrices and presented a similar formulation. The principal differences in this work are that we are using multilingual subspace parameters as priors, and we have applied it in a cross-lingual setting. In addition, we also note that it is possible to estimate the covariances using a data driven approach. For instance, a fully Bayesian treatment [8] can be applied, by which covariances can be estimated by maximizing the marginal likelihood

$$\arg\max_{\mathbf{\Omega}_r,\mathbf{\Omega}_c}\sum_i\int P(\mathbf{O}|\mathbf{M}_i)P(\mathbf{M}_i|\bar{\mathbf{M}}_i,\mathbf{\Omega}_r,\mathbf{\Omega}_c)d\mathbf{M}_i, \tag{10}$$

where $\mathbf{O}$ denotes all the acoustic frames. The likelihood $P(\mathbf{O}|\mathbf{M}_i)$ can be approximated by its lower bound, i.e. the auxiliary function (4), and as we use the conjugate prior to the auxiliary function, the analytical form of the marginal likelihood is available. Hence, this approach is expected to be feasible in practice. We have not experimentally investigated this approach in this paper.

3.3. MAP Adaptation of the Phonetic Subspace

The detailed analytical solution of the MAP estimate of subspace parameters with Gaussian prior is given by Povey [7] (App. J). Here, we summarize the main ideas. By substituting (4) and (7) into (6), the auxiliary function of MAP can be rewritten as:

$$\tilde{\mathcal{Q}}(\mathbf{M}_i) \propto tr\Big(\mathbf{M}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{Y}_i + \tau\mathbf{M}_i^T\mathbf{\Omega}_r^{-1}\bar{\mathbf{M}}_i\mathbf{\Omega}_c^{-1}\Big) - \frac{1}{2}tr\Big(\mathbf{\Sigma}_i^{-1}\mathbf{M}_i\mathbf{Q}_i\mathbf{M}_i^T + \tau\mathbf{\Omega}_r^{-1}\mathbf{M}_i\mathbf{\Omega}_c^{-1}\mathbf{M}_i^T\Big). \tag{11}$$

The solution is not readily available by taking the derivative of $\tilde{\mathcal{Q}}(\mathbf{M}_i)$ with respect to $\mathbf{M}_i$ and setting it to be zero. Instead, we introduce an intermediate transform $\mathbf{T} = \mathbf{U}^T\mathbf{L}^{-1}$ that simultaneously diagonalises $\mathbf{\Sigma}_i^{-1}$ and $\mathbf{\Omega}_r^{-1}$, where

$$\mathbf{\Sigma}_i^{-1} = \mathbf{L}\mathbf{L}^T \quad \text{(Cholesky decomposition)}, \tag{12}$$

$$\mathbf{S} = \mathbf{L}^{-1}\mathbf{\Omega}_r^{-1}\mathbf{L}^{-T}, \tag{13}$$

$$\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^T \quad \text{(Eigenvalue decomposition)}. \tag{14}$$

It is the case that $\mathbf{T}\mathbf{\Sigma}_i^{-1}\mathbf{T} = \mathbf{I}$ and $\mathbf{T}\mathbf{\Omega}_r^{-1}\mathbf{T} = \mathbf{\Lambda}$, where $\mathbf{I}$ is the identity matrix, and $\mathbf{\Lambda}$ is a diagonal matrix holding the eigenvalues

Table 1. WER (%) of baseline monolingual and cross-lingual baseline systems with 1 hour and 5 hour training data, S denotes the dimension of phonetic subspace.

1 hour training data	WER	#states	# sub-states
Mono-GMM	41.2	620	-
Mono-SGMM $S = 20$	38.0	620	2k
Cross-SGMM $S = 20$	35.0	620	12.8k
Cross-SGMM $S = 40$	32.7	620	4.4k
5 hour training data	WER	#states	#sub-states
Mono-GMM	34.3	1561	-
Mono-SGMM $S = 20$	31.1	1561	6.7k
Cross-SGMM $S = 20$	28.6	1561	12k
Cross-SGMM $S = 40$	26.8	1561	12k

of matrix $\mathbf{S}$. If we further define $\mathbf{M}_i = \mathbf{T}^T\mathbf{M}_i^{'}$, then equation (11) can be rewritten as

$$\tilde{\mathcal{Q}}(\mathbf{M}_i^{'}) \propto tr\Big(\mathbf{M}_i^{'T}\mathbf{T}\big(\mathbf{\Sigma}_i^{-1}\mathbf{Y}_i + \tau\mathbf{\Omega}_r^{-1}\bar{\mathbf{M}}_i\mathbf{\Omega}_c^{-1}\big)\Big) - \frac{1}{2}tr\Big(\mathbf{M}_i^{'}\mathbf{Q}_i\mathbf{M}_i^{'T} + \tau\mathbf{\Lambda}\mathbf{M}_i^{'}\mathbf{\Omega}_c^{-1}\mathbf{M}_i^{'T}\Big). \quad (15)$$

Now we can take the derivative of $\tilde{\mathcal{Q}}(\mathbf{M}_i^{'})$ with respect to $\mathbf{M}_i^{'}$:

$$\frac{\partial\tilde{\mathcal{Q}}(\mathbf{M}_i^{'})}{\partial\mathbf{M}_i^{'}} = \mathbf{T}\big(\mathbf{\Sigma}_i^{-1}\mathbf{Y}_i + \tau\mathbf{\Omega}_r^{-1}\bar{\mathbf{M}}_i\mathbf{\Omega}_c^{-1}\big) - \mathbf{M}_i^{'}\mathbf{Q}_i - \tau\mathbf{\Lambda}\mathbf{M}_i^{'}\mathbf{\Omega}_c^{-1}.$$

Setting this derivative to be zero, we obtain the row by row solution of $\mathbf{M}_i^{'}$ as

$$\mathbf{m}_n^{'} = \mathbf{g}_n\big(\mathbf{Q}_i + \tau\lambda_n\mathbf{\Omega}_c^{-1}\big)^{-1}, \quad (16)$$

where $\mathbf{m}_n^{'}$ is the n_{th} row of $\mathbf{M}_i^{'}$, λ_n is the n_{th} diagonal element of $\mathbf{\Lambda}$, and $\mathbf{g}_n$ is the n_{th} row of matrix $\mathbf{T}\big(\mathbf{\Sigma}_i^{-1}\mathbf{Y}_i + \tau\mathbf{\Omega}_r^{-1}\bar{\mathbf{M}}_i\mathbf{\Omega}_c^{-1}\big)$. The final solution of $\mathbf{M}_i$ can then be obtained by $\mathbf{M}_i = \mathbf{T}^T\mathbf{M}_i^{'}$.

As noted before, by setting $\tau \to \infty$, we shrink the system back to the cross-lingual baseline, and $\tau = 0$ corresponds to the ML estimate. If $\mathbf{\Omega}_r = \mathbf{\Omega}_c = \mathbf{I}$, the MAP estimate is equivalent to applying ℓ_2-norm regularization on $\mathbf{M}_i$ with the model origin set to be the multilingual estimate (cf. equation (11)). In this paper, we also study the roles of $\mathbf{\Omega}_r$ and $\mathbf{\Omega}_c$ individually, by setting the other to be $\mathbf{I}$.

4. EXPERIMENTAL RESULTS

We have carried out experiments using a cross-lingual acoustic model trained using the GlobalPhone corpus [9]. We chose German to be the target language, and Spanish, Portuguese and Swedish as source languages. In low-resource cross-lingual experiments, we selected two random subsets of transcribed audio in the target language, of 1 hour and 5 hours duration, containing speech from 8 and 40 speakers respectively[1]. We estimated the globally shared parameters in a multilingual fashion by tying $\mathbf{M}_i$, $\mathbf{w}_i$, and $\mathbf{\Sigma}_i$ across the three source language SGMM systems. The number of Gaussians I was 400 (cf. equation (1)). The models were evaluated on a development data set which containing about 2 hours of speech. For decoding, we used a trigram language model with a 17,000 word lexicon that was provided with the corpus. The language model (LM) had a perplexity of 442 on the development set, with an out of vocabulary (OOV) rate of 5.2%. Further details of this cross-lingual system can be found in [3].

[1]Although German should not be considered a low resource language, the GlobalPhone corpus provides a controlled and standardised experimental environment for experiments of this nature.

Table 2. WER (%) of MAP adapted systems.

System	1 hour	5 hour
Cross-lingual baseline($S = 20$)	35.0	28.6
with ML update ($\tau = 0$)	33.5	27.7
with MAP update ($\mathbf{I} \otimes \mathbf{I}$)	32.1	26.7
with MAP update ($\mathbf{I} \otimes \mathbf{\Omega}_c$)	32.2	26.8
with MAP update ($\mathbf{\Omega}_r \otimes \mathbf{I}$)	32.2	26.8
with MAP update ($\mathbf{\Omega}_r \otimes \mathbf{\Omega}_c$)	32.2	26.9
Cross-lingual baseline($S = 40$)	32.7	26.8
with ML update ($\tau = 0$)	33.3	27.8
with MAP update ($\mathbf{I} \otimes \mathbf{I}$)	31.1	25.6
with MAP update ($\mathbf{I} \otimes \mathbf{\Omega}_c$)	31.3	25.9
with MAP update ($\mathbf{\Omega}_r \otimes \mathbf{I}$)	31.1	25.5
with MAP update ($\mathbf{\Omega}_r \otimes \mathbf{\Omega}_c$)	31.4	25.8

4.1. Baseline results

The results of monolingual and cross-lingual German systems, with different amounts of training data and sizes of phonetic subspace, are given in Table 1. In the monolingual systems, all the parameters are estimated from the 1 or 5 hours of available training data. In cross-lingual SGMM systems, the globally shared parameters are taken from a multilingual system trained on Spanish, Portuguese and Swedish, and only sub-state vectors $\mathbf{v}_{jm}$ and weights c_{jm} (equation 1, 2) are updated during model training. The GMM and SGMM systems for the same amount of training data use the same phonetic decision tree. Hence, the performance differences are purely owing to better parameter estimation. For SGMM systems with $S = 40$, regularized state vector estimation by ℓ_1-norm penalty [10] is applied to improve numerical stability; we have also observed that such regularisation brings gains in accuracy [3]. For comparison, the monolingual GMM and SGMM systems with the entire 14.8 hours of target language training data available in GlobalPhone achieve 25.7% and 24.0% WER[2].

4.2. MAP adaptation results

Our MAP experiments started from the cross-lingual SGMM systems in Table 1, with the MAP update of $\mathbf{M}_i$ performed for several iterations until convergence, while $\mathbf{w}_i$ and $\mathbf{\Sigma}_i$ were kept fixed. We compare different configurations of the row and column covariances for the priors as shown in Table 2, where the results are obtained by tuning the smoothing parameter τ to be optimal on the development set. As mentioned before, setting $\tau = 0$ is equivalent to an ML update of $\mathbf{M}_i$. When $S = 20$, ML update provided considerable improvements with both the 1 hour and 5 hour data, since the number of parameters to be updated is relatively small. But for systems with $S = 40$, which have a much larger number of parameters, we observed an increase in WER. MAP update, on the other hand, provided consistent reductions in WER.

For systems with $S = 20$ MAP update gave an additional 1% absolute WER reduction compared with ML update, in both training conditions. For their counterparts with $S = 40$, MAP update resulted in 1% absolute reduction in WER compared with the baseline, whereas the ML update increased the WER. This is consistent with our expectation that MAP can overcome the model overfitting encountered by ML. Again we used ℓ_1-norm regularized (sub-)state vector estimation [10] to improve numerical stability. However, we do not observe any improvement in WER by using full row and column covariance matrices compared to the identity matrices used in

[2]The weak LM and lexicon of high OOV rate lead to relatively poor baseline systems.

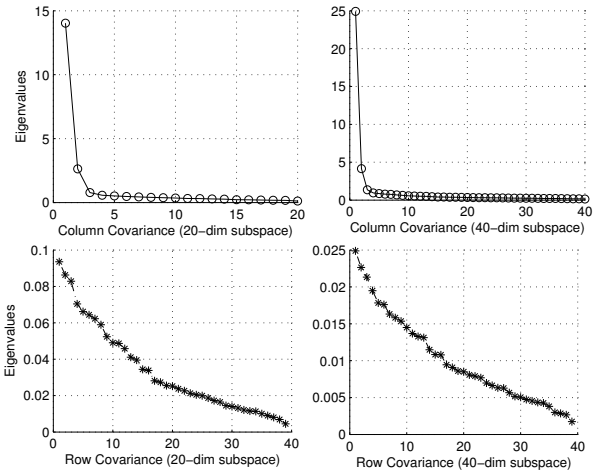

Fig. 1. Eigenvalues of row and column covariance matrices $\mathbf{\Omega}_r$, $\mathbf{\Omega}_c$ for both 20 and 40 dimensional subspace.

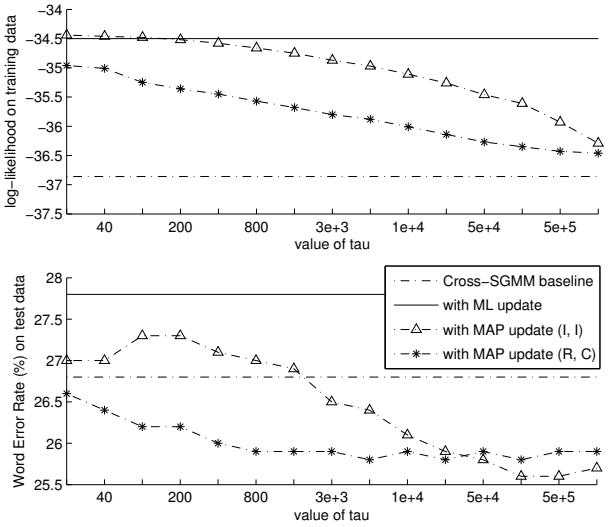

Fig. 2. Effect of smoothing parameter τ in MAP adapted system on the log-likelihood in the training stage and WER in the testing stage.

the priors. Setting just one of two covariance matrices to be the identity, did not result in a significant difference.

To obtain a better understanding of these results, we have plotted the eigenvalues of $\mathbf{\Omega}_r$ and $\mathbf{\Omega}_c$ for systems with $S = 20$ and $S = 40$ (Figure 1). This shows that the eigenvalues of column covariances decrease rapidly, and that the first few eigenvectors corresponding to the top eigenvalues account for most of the variance. This was unexpected, as priors with such a covariance structure will constrain the model to model subspace of lower effective dimension, and limit its ability to learn from the data. In addition, the ad hoc approach we have employed to approximate the covariances of the prior may not be optimal. We cannot guarantee that the global covariance from the multilingual subspace will work well for the target system. In future work, we shall investigate the estimation of $\mathbf{\Omega}_r$, $\mathbf{\Omega}_c$ using the Bayesian approach of equation (10).

Finally, figure 2 shows the effect of the smoothing parameter τ for both training and testing for a system with $S = 40$ and 5 hours of training data. Here, we only show the MAP systems with row and column covariance matrices in the priors to be both identity or full, denoted as "(I, I)" and "(R, C)", respectively. When τ is small, the log-likelihood is close to that of ML system, and as τ increases, the log-likelihood decreases accordingly, but it is bounded by the baseline system which corresponds to $\tau \to \infty$. On the other hand, by tuning the value of τ, the WER of a MAP adapted system can be smaller than both baseline and ML system. Other MAP adapted systems in Table 2 show a similar trend. Note that the absolute value of optimal τ depends on the prior distribution and also the amount of training data which means its range varies for different systems.

5. CONCLUSION

In this paper, we investigated the MAP adaptation of the phonetic subspace parameters in an SGMM acoustic model for cross-lingual speech recognition. In this approach, a matrix variate Gaussian prior is introduced to the subspace parameter estimation in order to avoid model overfitting in limited resource conditions. In our cross-lingual speech recognition experiments the phonetic subspace parameters estimated in the multilingual system served as priors for the target language systems. Experiments on the GlobalPhone corpus indicated that considerable reductions in WER are given by this MAP adaptation approach. In future work, we plan to apply the MAP adaptation algorithm presented in this paper to the speaker subspace of an SGMM acoustic model in a cross-lingual setting. In addition, a Bayesian estimation of the prior parameters will be experimentally explored, as well as the adaptation of the weight projections.

6. REFERENCES

[1] D. Povey, L. Burget, et al., "The subspace Gaussian mixture model—A structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[2] L. Burget, P. Schwarz, et al., "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *Proc. IEEE ICASSP*, 2010, pp. 4334–4337.

[3] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian mixture models for cross-lingual speech recognition," in *Proc. IEEE ASRU*, 2011.

[4] A.K. Gupta and D.K. Nagar, *Matrix Variate Distributions*, vol. 104 of *Monographs and Surveys in Pure and Applied Mathematics*, Chapman & Hall/CRC, 1999.

[5] O. Siohan, C. Chesta, and C.H. Lee, "Joint maximum a posteriori adaptation of transformation and HMM parameters," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 417–428, 2001.

[6] P. Dutilleul, "The MLE algorithm for the matrix normal distribution," *Journal of Statistical Computation and Simulation*, vol. 64, no. 2, pp. 105–123, 1999.

[7] D. Povey, "A tutorial-style introduction to subspace Gaussian mixture models for speech recognition," Tech. Rep., MSR-TR-2009-111, Microsoft Research, 2009.

[8] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer New York:, 2006.

[9] T. Schultz, "GlobalPhone: a multilingual speech and text database developed at Karlsruhe University," in *Proc. ICLSP*, 2002, pp. 345–348.

[10] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian mixture models for speech recognition," *IEEE Signal Processing Letters*, vol. 18, no. 7, pp. 419–422, 2011.