CONSTRAINED MLE-BASED SPEAKER ADAPTATION WITH L1 REGULARIZATION

Younggwan Kim, Hoirin Kim

Department of Electrical Engineering, KAIST, Daejeon 305-701, Korea

ABSTRACT

Maximum *a posterior* (MAP) adaptation is one of the popular and powerful methods for obtaining a speaker-specific acoustic model. Basically, MAP adaptation needs a data storage for speaker adaptive (SA) model as much as speaker independent (SI) model needs. Modern speech recognition systems have a huge number of parameters and deal with millions of users. To reduce the data storage for SA models, in this paper, we propose a constrained maximum likelihood estimation-based speaker adaptation with L1 regularization. By the proposed method, we can more efficiently perform the model adjustments for SA models without almost any loss of phone recognition performance than the conventional sparse MAP adaptation method.

Index Terms— Speaker adaptation, maximum *a posterior* adaptation, constrained MLE, L1 regularization, Euclidean projection on L1 ball

1. INTRODUCTION

Current speech recognition systems using hidden Markov models (HMMs) have employed speaker adaptation methods to improve robustness against speaker variability. There have been various adaptation techniques such as maximum likelihood linear regression (MLLR) [1, 2], eigenvoice (EV) adaptation [3, 4], and maximum *a posteriori* (MAP) adaptation [5]. Typically, MLLR and EV-based methods are well known adaptation techniques for very limited adaptation data (10 seconds to 10 minutes) and require small amount of speaker-specific parameters compared with a speaker independent (SI) model. On the contrary, it has been known that MAP adaptation is good for medium amount of adaptation data (20 minutes to 10 hours) and requires a number of parameters as much as a SI model has.

In [5], MAP adaptation employs Bayesian priors for the Gaussian components and the priors have some effects same as L2 norm regularization. Typically, L2 norm regularization may cause many small parameter adjustments in the adaptation processing. Olsen *et al*. [6, 7] showed that most of the adapted parameters are not closely related to speech recognition performance. So they proposed sparse MAP adaptation for limiting the redundant parameter adjustments. In order to obtain the best results in the sparse MAP adaptation, they controlled several parameters which are related to hyperparameters for adaptation and Lagrangian multipliers for sparsity. As the number of parameters increases, it becomes hard to find optimal parameter values for the best performance of recognition systems. To cope with this difficulty, we reinterpret the MAP adaptation as a constrained optimization problem and propose a constrained maximum likelihood estimation (CMLE)-based speaker adaptation method with L1 regularization. The proposed method is a MAP-like adaptation algorithm which considers simultaneously controlling both regularization and sparsity and, converges to maximum likelihood estimation as the amount of adaptation data increases.

2. GEOMETRIC REVIEW OF MAP ADPATATION

The basic approach of obtaining a speaker-adaptive (SA) acoustic model is to derive the SA model by adapting a speaker independent (SI) model [5]. In order to adapt the SI model, the GMM-based MAP adaptation process is well described in [8]. Sufficient statistics computed by the maximum likelihood criterion are used to obtain the SA model as follows.

$$w_i^{\mathrm{MAP}} = (\alpha_i^w S_{w,i} + (1-\alpha_i^w) w_i^{\mathrm{SI}}) \big/ \eta, \tag{1}$$

$$\boldsymbol{\mu}_i^{\mathrm{MAP}} = \alpha_i^{\mu} \mathbf{S}_{\boldsymbol{\mu},i} + (1-\alpha_i^{\mu}) \boldsymbol{\mu}_i^{\mathrm{SI}}, \tag{2}$$

$$\mathbf{v}_i^{\mathrm{MAP}} = \alpha_i^{v} \mathbf{S}_{\mathbf{v},i} + (1-\alpha_i^{v})(\mathbf{v}_i^{\mathrm{SI}} + (\boldsymbol{\mu}_i^{\mathrm{SI}})^2) - (\boldsymbol{\mu}_i^{\mathrm{MAP}})^2, \tag{3}$$

where α^{κ}, $\kappa \in \{w, \mu, v\}$, is the adaptation coefficients for the weights, means, and variances, respectively, η is a scale factor which enables summation of the adapted mixture weight to be unity, $\mathbf{S}_{\theta,i}$ indicates the sufficient statistics for each Gaussian model parameter, and $\mathbf{v}_i$ is the variance vector which constitutes diagonal components of covariance matrix. The adaptation coefficient α^{κ} is used to determine the balance between the sufficient statistics and SI model parameters. The adaptation coefficient is given $\alpha_i^{\kappa} = n_i \big/ \left(n_i + \tau^{\kappa}\right)$, where n_i is the posterior sum of mixture i and τ^{κ} is a hyperparameter.

As can be seen in equation (1)-(3), it is noticeable that all adapted model parameters are computed by the same

form of interpolation and the interpolation operation can be treated as a regularization of MAP adaptation. From this point of view, MAP adaptation can be regarded as a constrained optimization problem [9], such that

$$\begin{aligned} &\min_{\boldsymbol{\varphi}_i} \frac{1}{2}\left\|\boldsymbol{\varphi}_i - \left(\mathbf{S}_i - \boldsymbol{\theta}_i^{\mathrm{SI}}\right)\right\|_2^2 \\ &\text{subject to } \left\|\boldsymbol{\varphi}_i\right\|_2 \le \left\|\left(\mathbf{S}_i - \boldsymbol{\theta}_i^{\mathrm{SI}}\right)\frac{n_i}{n_i + \tau^{\kappa}}\right\|_2, \end{aligned} \tag{4}$$

where $\boldsymbol{\varphi}_i$ denotes a vector moving from SI model, $\mathbf{S}_i$ and $\boldsymbol{\theta}_i^{\mathrm{SI}}$ represents sufficient statistics and SI models, and $\|\cdot\|_2$ is L2 norm. In the optimization problem, note that $\mathbf{S}_i$ and $\boldsymbol{\theta}_i^{\mathrm{SI}}$ are used for shared notations of weight, mean, and variance. Finally, we get the adapted model parameter given by

$$\boldsymbol{\theta}_i^{\mathrm{MAP}} = \boldsymbol{\varphi}_i^{\mathrm{MAP}} + \boldsymbol{\theta}_i^{\mathrm{SI}} = \frac{n_i \mathbf{S}_i + \tau^{\kappa}\boldsymbol{\theta}_i^{\mathrm{SI}}}{n_i + \tau^{\kappa}}. \tag{5}$$

In (5), we ignore the scaling factor in (1) and the vector subtraction of squared adapted mean vector in (3). In (1), the equation for adapted weight is defined in 1-dimensional space. Since, however, we can also treat the weight components as a vector form in a Gaussian mixture model, it is also possible that the weight vector interpolation also can be interpreted as the constrained optimization problem. This constrained optimization problem is described in Fig. 1 from a geometrical perspective. As can be seen in Fig. 1, the shaded region implies constraint part of the optimization problem and the interpolation form of (1)-(3) is totally caused by L2 norm-based constraint. This is the reason why MAP adaptation is also called L2 regularization and the L2 norm-based constraint causes most of the small and redundant adjustments which can be negligible in terms of the speech recognition performance.

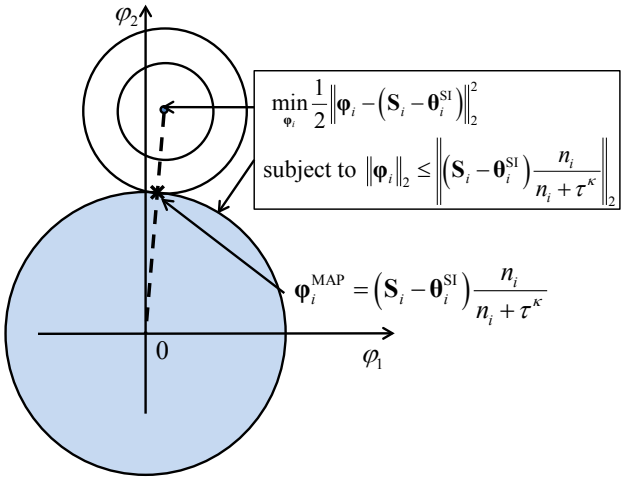

Fig. 1. Geometric interpretation for MAP adaptation

3. SPARSE MAP ADAPTATION

In [6], Olsen *et al*. showed that most of the adapted parameters, whose differences between the SA and the SI models were quite small after MAP adaptation, were irrelevant to speech recognition performance. It was also mentioned that, during adaptation, ignoring the small adjustment could be helpful to save on data storage for the SA models and could also improve recognition performance. For these reasons, the adapted model parameters are selected by constrained optimization problem in sparse MAP adaptation. The constrained optimization problem is

$$\begin{aligned} &\max_{\boldsymbol{\Theta}} \sum_{i=1}^{G}\sum_{d=1}^{D}(n_i + \tau^{\xi})L(\mathbf{X};\boldsymbol{\theta}_{i,d} \in \{\boldsymbol{\theta}_{i,d}^{\mathrm{MAP}}, \boldsymbol{\theta}_{i,d}^{\mathrm{SI}}\}) \\ &\text{subject to } N = \sum_{i=1}^{G}\sum_{d=1}^{D}\left\|\boldsymbol{\theta}_{i,d} - \boldsymbol{\theta}_{i,d}^{\mathrm{SI}}\right\|_0, \end{aligned} \tag{6}$$

where $\boldsymbol{\Theta}$ indicates the total parameter set of acoustic model, G is the number of Gaussians, D is the dimension of the acoustic feature vector, $\boldsymbol{\theta}_{i,d}=\{\boldsymbol{\mu}_{i,d}, \mathbf{v}_{i,d}\}$ is a set of parameters for dimension d of Gaussian i, and $\|\boldsymbol{\theta}_{i,d} - \boldsymbol{\theta}_{i,d}^{\mathrm{SI}}\|_0 \in \{0, 1, 2\}$ is L0 norm. $L(\cdot)$ is the log likelihood function that calculates likelihood of feature vectors with given model parameters. Comparing (4) and (6), the optimization problem in (6) allows only N parameters to be changed by the constraint. Since mixture weight is not considered in this adaptation method, hyperparameter set consists of $\xi \in \{\mu, \nu\}$. Since the objective function and constraints in (6) are composed of direct sums over i and d, the problem can be solved exactly. The optimal solution for an N can be obtained by a greedy search for N parameters that most increase the objective function. For example, when $N = 1$, it is clear that the single MAP-adapted parameter causing the largest increase in the objective function is selected for the SA model and all other remaining parameters take SI model parameters.

From (6), we can alternatively consider maximizing the Lagrangian function given by

$$\begin{aligned} \mathcal{L}_{\mathrm{SMAP}}(\boldsymbol{\Theta};\lambda) = &\sum_{i=1}^{G}\sum_{d=1}^{D}(n_i + \tau^{\xi})L(\mathbf{X};\boldsymbol{\theta}_{i,d} \in \{\boldsymbol{\theta}_{i,d}^{\mathrm{MAP}}, \boldsymbol{\theta}_{i,d}^{\mathrm{SI}}\}) \\ &- \lambda\Big(\sum_{i=1}^{G}\sum_{d=1}^{D}\left\|\boldsymbol{\theta}_{i,d} - \boldsymbol{\theta}_{i,d}^{\mathrm{SI}}\right\|_0 - N\Big), \end{aligned} \tag{7}$$

where λ is a Lagrangian multiplier. As mentioned in [6], for fixed λ, (7) can be fully decoupled across i, d and thus each sub-problem is given as follows:

$$\max_{\boldsymbol{\theta}_{i,d}} (n_i + \tau^{\xi})L(\mathbf{X};\boldsymbol{\theta}_{i,d} \in \{\boldsymbol{\theta}_{i,d}^{\mathrm{MAP}}, \boldsymbol{\theta}_{i,d}^{\mathrm{SI}}\}) - \lambda\left\|\boldsymbol{\theta}_{i,d} - \boldsymbol{\theta}_{i,d}^{\mathrm{SI}}\right\|_0. \tag{8}$$

For computational benefit, it is useful to minimize the following equation

$$F(\boldsymbol{\theta}_{i,d};\alpha) = -2L(\mathbf{X};\boldsymbol{\theta}_{i,d} \in \{\boldsymbol{\theta}_{i,d}^{\text{MAP}}, \boldsymbol{\theta}_{i,d}^{\text{SI}}\}) + \alpha \left\| \boldsymbol{\theta}_{i,d} - \boldsymbol{\theta}_{i,d}^{\text{SI}} \right\|_0, \tag{9}$$

$$\alpha = \frac{2\lambda}{n_i + \tau^{\xi}}. \tag{10}$$

Besides L0 norm in (9), L1 norm also can be applied to the optimization problem. Similarly, in [7], new methods using the combination of L0 and L1 norm for the optimization problem were also proposed.

4. PROPOSED METHOD

Sparse MAP adaptation needs to control additional Lagrangian multipliers as well as hyperparameters. It may cause difficulties in finding optimal values showing the best recognition performance. To obtain the sparsity and regularization effect simultaneously, we were inspired to apply L1 norm-based constraint from interpreting the MAP adaptation as a constrained optimization problem. Based on the idea, CMLE-based speaker adaptation with L1 regularization is proposed in this paper. The proposed optimization problem is given as follows

$$\min_{\boldsymbol{\varphi}_i} \frac{1}{2}\left\| \boldsymbol{\varphi}_i - \left(\mathbf{S}_i - \boldsymbol{\theta}_i^{\text{SI}}\right) \right\|_2^2 \quad \text{subject to } \left\| \boldsymbol{\varphi}_i \right\|_1 \le \left\| \left(\mathbf{S}_i - \boldsymbol{\theta}_i^{\text{SI}}\right) \frac{n_i}{n_i + \tau^{\kappa}} \right\|_2, \tag{11}$$

where $\|\cdot\|_1$ is L1 norm. As can be seen in (11), the idea was mainly based on (4) from which we changed the L2 norm-based constraint into the L1 norm-based constraint. This type of constrained optimization problem is typically called the problem of Euclidean projection onto the L1 ball [10].

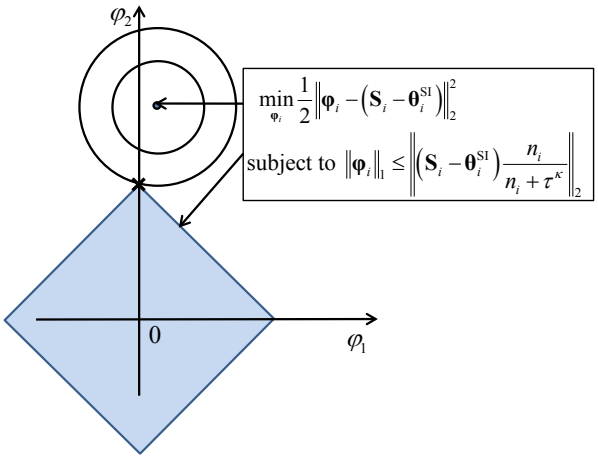

Fig. 2. Geometric interpretation for CMLE with L1 regularization

By writing the Lagrangian form of (11), we have

$$\mathcal{L}_{\text{CMLE-L1}}(\boldsymbol{\varphi}_i;\lambda) = \frac{1}{2}\left\| \boldsymbol{\varphi}_i - \left(\mathbf{S}_i - \boldsymbol{\theta}_i^{\text{SI}}\right) \right\|_2^2 + \lambda \left(\left\| \boldsymbol{\varphi}_i \right\|_1 - \left\| \left(\mathbf{S}_i - \boldsymbol{\theta}_i^{\text{SI}}\right) \frac{n_i}{n_i + \tau^{\kappa}} \right\|_2 \right). \tag{12}$$

Compared with sparse MAP adaptation, strong duality holds for (11), and the primal and dual optimal values are equal. Thus, we can determine the primal optimal points for model parameters by finding dual optimal point for Lagrangian multiplier. When the dual optimal point λ^* is known, we can find the primal solution by

$$\boldsymbol{\varphi}_i^{\text{CMLE-L1}} = \arg\min_{\boldsymbol{\varphi}_i} \mathcal{L}_{\text{CMLE-L1}}(\boldsymbol{\varphi}_i;\lambda^*). \tag{13}$$

Since each dimension can be fully decoupled into individual variable, we have

$$\varphi_{i,d}^{\text{CMLE-L1}} = \arg\min_{\varphi_{i,d}} \frac{1}{2}\left(\varphi_{i,d} - \left(S_{i,d} - \theta_{i,d}^{\text{SI}}\right)\right)^2 + \lambda^* \left(\left|\varphi_{i,d}\right| - \left\| \left(\mathbf{S}_i - \boldsymbol{\theta}_i^{\text{SI}}\right) \frac{n_i}{n_i + \tau^{\kappa}} \right\|_2 \right), \tag{14}$$

which leads to

$$\varphi_{i,d}^{\text{CMLE-L1}} = \text{sign}\left(S_{i,d} - \theta_{i,d}^{SI}\right) \max\left(\left|S_{i,d} - \theta_{i,d}^{SI}\right| - \lambda^*, 0\right) \tag{15}$$

$$\theta_{i,d}^{\text{CMLE-L1}} = \varphi_{i,d}^{\text{CMLE-L1}} + \theta_{i,d}^{SI}, \tag{16}$$

where sign(r) returns +1 if $r \ge 0$ and –1 otherwise. To find dual optimal point λ^*, the bisection algorithm is typically used and the detailed procedure for obtaining λ^* is well described in [10].

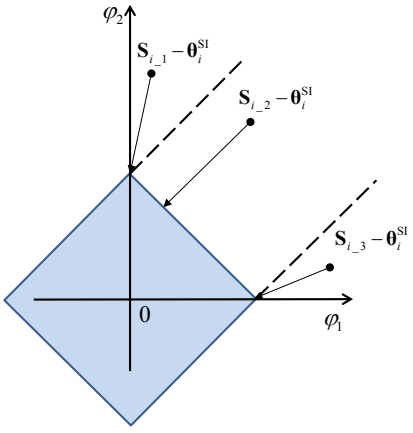

Fig. 3. Illustration of the Euclidean projection onto the L1 ball

With the proposed method, we can simultaneously control the sparsity and regularization effect by choosing hyperparameters τ^K. Furthermore, our method is a more generalized sparse adaptation method since every model parameter including mixture weights can be adapted in our constrained optimization problem. In Fig. 2, the effect of L1 norm-based constraint is illustrated by a geometrical perspective. The shaded region in Fig. 2 is the constraint part of the optimization problem (11) and it is also expanded by posterior sum n_i same as (4) in MAP adaptation. As can be seen in the figure, it is also observed that the L1 ball acts as a sparsity promoting regularizer. From the property, we can obtain the sparsely updated SA model. In addition, Fig. 3 shows how the Euclidean projection onto the L1 ball works with three different cases of sufficient statistics. From the Fig. 3, we can notice that the adapted model is very sparsely updated from SI model, when n_i is small in comparison with the hyperparameter τ^K and vice versa.

5. EXPERIMENTS

The experiments were performed on the ETRI Korean conversation speech database which had been collected at 16 kHz sampling rate and 16-bit resolution by two types of smart phone devices in clean environmental condition. This database contains a total of 52,500 sentences, which are composed of 150 sentences spoken by each of 350 speakers, and each sentence is composed of about 4 to 5 seconds long. We used 45,000 sentences of 300 speakers for constructing the SI triphone models and remaining speech data of 50 speakers for test. We used 13 dimensional Mel-frequency cepstral coefficients and their first and seconds derivatives as a feature vector. For adaptation, we used 100 utterances per speaker and remaining 50 utterances were used for the phone recognition test. We applied phone level unigram language model of 39 Korean phonemes for our experiments. The SI model had 11,848 tied-state triphone HMMs including 3 states per each HMM and 32 Gaussian mixtures per state. For each SA model, we only adapted the mean vector from the SI model.

All tests are performed according to various values of hyperparameter τ. In order to solve the proposed optimization problem, we used SLEP toolbox in [11]. In Table 1, SMAP and CMLE-L1 indicate the results of the sparse MAP adaptation and proposed method, respectively, and phone error rate (PER) and sparsity of each method are summarized. For the comparison, we did our experiments on SI model and MLLR adaptation. For MLLR, we used 40 regression classes which also showed the best PER results from the same set of adaptation and test data. For SMAP, we used L0 norm based algorithm and $\lambda = 0.005$ in (10) which also showed the best performance for SMAP. From our experimental results, the proposed method can keep more model parameters unchanged (91.21% (SMAP) → 95.28% (CMLE-L1) in sparsity) compared with SMAP while maintaining comparable recognition performance in PER compared with the results of MAP. This indicates that only 4.72% of total model parameters are the key parameters influencing recognition performance.

Table 1. Phone error rate and sparsity for MAP, sparse MAP and proposed method

τ	2	1.5	1.2	1	0.5
Phone error rate (PER)					
SI Model	31.45%				
MLLR	25.81%				
MAP	22.58%	22.25%	**22.04%**	22.68%	23.33%
SMAP	22.67%	22.38%	**22.23%**	22.74%	23.20%
CMLE-L1	22.10%	22.36%	22.58%	**22.06%**	23.42%
Sparsity					
MAP	**70.19%**	**70.19%**	**70.19%**	**70.19%**	**70.19%**
SMAP	93.59%	92.64%	**91.21%**	90.33%	88.24%
CMLE-L1	96.67%	95.98%	95.61%	**95.28%**	93.45%

6. CONCLUSION

In this paper, MAP adaptation was reinterpreted as a constrained optimization problem with L2 regularization. From this point of view, we proposed CMLE-based speaker adaptation methods with L1 regularization. By the proposed method, it is observed that we can achieve more sparsity than L0 norm-based SMAP without almost any loss of PER compared with MAP. In addition, it is also shown that we can control both sparsity and regularization only by adjusting the hyperparameter without any additional parameters. Further work is to finish the tests with adapted weights and variances and to find combinatorial constraint which can improve the recognition performance and sparsity more.

7. ACKNOWLEDGEMENT

This work was supported by the IT R&D program of MSIP/KEIT. [10041807, Development of Original Software Technology for Automatic Speech Translation with Performance 90% for Tour/International Event focused on Multilingual Expansibility and based on Knowledge Learning]

8. REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMM's," *Comput. Speech Lang*., vol. 9, pp. 171–186, 1995.

[2] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," CUED/F-INFENG Technical Report 291, Cambridge Univ. Eng. Dept., 1997.

[3] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[4] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process*, vol. 13, no. 3, pp. 345–354, May 2005.

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

[7] P. A. Olsen, J. Huang, S. J. Rennie, and V. Goel, "Affine invariant sparse maximum a posteriori adaptation," in *Proc. ICASSP*, 2012, pp. 4317 – 4320.

[8] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometric Recognition*. Cambridge, MA: MIT Lincoln Laboratory, 2008, pp. 12–17.

[9] C.M. Bishop, *Pattern recognition and machine learning*, Information Science and Statistics. Springer-Verlag, 2nd edition, 2006.

[10] J. Liu and J. Ye, "Efficient Euclidean projections in linear time," in *Proc.26th Annu. Int. Conf. Mach. Learn.*, Montreal, QC, Canada, 2009, pp.1–8.

[11] J. Liu, S. Ji, and J. Ye. SLEP: Sparse learning with efficient projections. Arizona State University, 2010, http://www.public.asu.edu/~jye02/Software/SLEP.