ONLINE SPEAKER ADAPTATION BASED ON QUASI-BAYES LINEAR REGRESSION

Jen-Tzung Chien and Chih-Hsien Huang

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
jtchien@mail.ncku.edu.tw

ABSTRACT

This paper presents an online/sequential linear regression adaptation framework for hidden Markov model (HMM) based speech recognition. Our attempt is to sequentially improve speaker-independent (SI) speech recognizer to meet nonstationary environments via linear regression adaptation of SI HMM's. A *quasi-Bayes linear regression* (QBLR) algorithm is developed to execute online adaptation where the regression matrix is estimated using QB theory. In the estimation, we moderately specify the prior density of regression matrix as a *matrix variate normal distribution* and exactly derive the pooled posterior density belonging to the same distribution family. Accordingly, the optimal regression matrix can be easily calculated. Also, the reproducible prior/posterior density pair provides meaningful mechanism for sequential learning of prior statistics. At each sequential epoch, only the updated prior statistics and the current observed data are required for adaptation. In general, the proposed QBLR is universal and can be reduced to well-known maximum likelihood linear regression (MLLR) and maximum *a posteriori* linear regression (MAPLR). Experiments show that the QBLR is effective for speaker adaptation in car environments.

1. INTRODUCTION

There is no doubt that the robustness issue is crucial for speech recognition in real-world applications because the mismatch between training and testing data always exists and degrades the performance considerably. One effective approach is to adjust the existing speech hidden Markov models (HMM's) to fit the acoustics of test speaker/noise/transducer. The desirable recognition performance can be obtained. However, because the realistic environments are time-variant, an attractive adaptation strategy is to sequentially perform the adaptation using online observed data instead of doing batch adaptation [2][9]. In the literature, the maximum likelihood linear regression (MLLR) [10] worked quite well for batch speaker adaptation. Using MLLR, the adaptation of $d \times 1$ HMM mean vector μ_{jk} of state *j* and mixture component *k* is achieved by applying a $d \times (d+1)$ cluster-dependent transformation matrix W_c to the extended mean vector $\xi_{jk} = [1, \mu_{jk}^T]^T$. The likelihood of observation $\mathbf{x}_t$ associated with the adapted mean vector becomes

$$p(\mathbf{x}_t | W_c, \lambda_{jk} = (\mu_{jk}, \Sigma_{jk})) = (2\pi)^{d/2} |\Sigma_{jk}|^{-1/2} \times \exp\left[-\frac{1}{2} (\mathbf{x}_t - W_c \xi_{jk})^T \Sigma_{jk}^{-1} (\mathbf{x}_t - W_c \xi_{jk}) \right], \quad (1)$$

where $\lambda = \{\lambda_{jk}\}$ are HMM parameters. The MLLR is aimed to estimate the set of regression matrices $W = \{W_c\}$ by maximizing the likelihood of batch adaptation data $\mathbf{X} = \{\mathbf{x}_t\}$

$$W_{ML} = \arg\max_W p(\mathbf{X} | W, \lambda) . \quad (2)$$

However, in case of sparse adaptation data, the ML estimation often leads to biased estimate. To deal with the sparseness problem, it is helpful to incorporate the prior knowledge and apply the maximum *a posteriori* (MAP) principle to estimate the regression matrix. The maximum *a posteriori* linear regression (MAPLR) [1][3][5] was accordingly established by

$$W_{MAP} = \arg\max_W p(W | \mathbf{X}, \lambda) \propto \arg\max_W p(\mathbf{X} | W, \lambda) g(W | \varphi) . \quad (3)$$

Generally, MAP estimate W_{MAP} could outperform ML estimate W_{ML} when the subjective prior density $g(W|\varphi)$ with hyper-parameters φ is proper.

In this study, our strategy is to develop an online linear regression adaptation scheme in case that the adaptation data $\chi^n = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ are sequentially observed. The mechanisms of *sequential parameter estimation* and *prior statistics evolution* should be formulated. In [2][9], the quasi-Bayes (QB) framework was exploited for online adaptation/transformation of HMM's. Herein, we would like to estimate the linear regression model parameter corresponding to *n*th observation epoch $W^{(n)}$ using the following QB principle

$$W_{QB}^{(n)} = \arg\max_W p(W | \chi^n, \lambda) \cong \arg\max_W p(\mathbf{X}_n | W, \lambda) g(W | \varphi^{(n-1)}) , \quad (4)$$

where the posterior density of overall data χ^n is approximated by a product of likelihood of current data $\mathbf{X}_n$ and prior density of regression matrix W given hyperparameters $\varphi^{(n-1)}$ evolved from history data χ^{n-1}. Starting from initial hyperparameters $\varphi^{(0)}$, the sequential learning of parameters $W^{(1)}, W^{(2)}, \cdots, W^{(n)}$ and evolution of hyperparameters $\varphi^{(1)}, \varphi^{(2)}, \cdots, \varphi^{(n)}$ can be established by continuously applying the adaptation data $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$. After refreshing hyperparameters $\varphi^{(n)}$, the current observation data $\mathbf{X}_n$ are thrown away at each sequential epoch. As a result, we don't need to store long batch data for adaptation so that the adaptation efficiency can be consolidated and the nonstationary environments can be continuously traced. In this paper, we construct and validate the general quasi-Bayes linear regression (QBLR) theory for online speaker adaptation in car speech recognition.

2. PRIOR DISTRIBUTION

There are two key issues for QB-based parameter estimation. First, the QB estimate in (4) should have a *close form solution* to achieve rapid linear regression adaptation. This issue is important because the calculation of a set of regression matrices using numerical approach, e.g. [11], would be very expensive. Second, the prior density of regression matrix should belong to the *conjugate distribution family* where the prior distribution and the pooled posterior distribution have the same distribution form regardless of the size and the values in observation data. With the conjugate prior specification, we could generate the reproducible prior/posterior density pair. The mechanism of hyperparameter evolution can be built for sequential adaptation. In [2], we used the cluster-dependent transformation function

$$\hat{\lambda} = \{\mu_{jk} + \mu_c^{(n)}, \theta_c^{(n)} r_{jk}\}, \tag{5}$$

where $r_{jk} = \Sigma_{jk}^{-1}$ is precision matrix. The joint prior density of transformation parameters $\mu_c^{(n)}$ and $\theta_c^{(n)}$ was defined by normal-Wishart density

$$\begin{aligned} g(\mu_c^{(n)}, \theta_c^{(n)} \big| \tau_c^{(n-1)}, m_c^{(n-1)}, \alpha_c^{(n-1)}, u_c^{(n-1)}) &\propto \left|\theta_c^{(n)}\right|^{(\alpha_c^{(n-1)}-d)/2} \\ &\times \exp\left[-\frac{1}{2}(\mu_c^{(n)} - m_c^{(n-1)})^T \theta_c^{(n)} \tau_c^{(n-1)} (\mu_c^{(n)} - m_c^{(n-1)})\right] \\ &\times \exp\left[-\frac{1}{2}\mathrm{tr}(u_c^{(n-1)}\theta_c^{(n)})\right]. \end{aligned} \tag{6}$$

It was proved that the resulting posterior density also had a normal-Wishart density form [2]. The online transformation of HMM's was proposed for supervised speaker adaptation. Herein, the QBLR is also an online transformation technique where the transformation function is constrained by a linear regression model. Therefore, the point of QBLR turns out to find a conjugate prior distribution for regression matrix.

Basically, the prior density should properly reflect the statistical behavior of parameter. However, the selection of prior density for matrix parameter is really tricky in Bayesian learning. As suggested in MAPLR [1][5], the $d \times (d+1)$ regression matrix of cluster c can be modeled by an elliptically symmetric distribution (or matrix variate normal distribution) [7][8]

$$\begin{aligned} g(W_c^{(n)} \big| \varphi_c^{(n-1)}) \propto \left|\Delta_c^{(n-1)}\right|^{-1/2} \exp\Bigg\{ \sum_{i=1}^{d} \left[W_c^{(n)}(i) - M_c^{(n-1)}(i)\right] \\ \times \Sigma_{ci}^{(n-1)^{-1}} \left[W_c^{(n)}(i) - M_c^{(n-1)}(i)\right]^T \Bigg\}, \end{aligned} \tag{7}$$

where $W_c^{(n)}(i)$ and $M_c^{(n-1)}(i)$ are respectively the *i*th rows of regression matrix $W_c^{(n)}$ and hyperparameter $M_c^{(n-1)}$ and the matrix $\Delta_c^{(n-1)} = diag(\Sigma_{c1}^{(n-1)}, \cdots, \Sigma_{cd}^{(n-1)})$ is a $d(d+1) \times d(d+1)$ block diagonal matrix with each diagonal block element $\Sigma_{ci}^{(n-1)}$ being $(d+1) \times (d+1)$ matrix. This distribution has location parameter $M_c^{(n-1)} = \{M_c^{(n-1)}(i)\}$ (mean vector) and scale matrix $\Delta_c^{(n-1)}$ (covariance matrix). Next, we will show the conjugate property on the use of matrix variate normal distribution. This property is crucial to develop QBLR algorithm for online speaker adaptation.

3. QUASI-BAYES LINEAR REGRESSION

To obtain QB estimate in (4), we apply EM algorithm [6] to overcome the missing data problem in HMM related formulation. Namely, the first step (E-step) is to calculate the expectation

$$\begin{aligned} R_{QB}(\hat{W}^{(n)} \big| W^{(n)}) = E\{\log p(\mathbf{X}_n, \mathbf{s}_n, \mathbf{l}_n \big| \hat{W}^{(n)}) \big| \mathbf{X}_n, W^{(n)}\} \\ + \log g(\hat{W}^{(n)} \big| \varphi^{(n-1)}) \end{aligned}, \tag{8}$$

where $W^{(n)}$ is current estimate, $\hat{W}^{(n)}$ is new estimate and $(\mathbf{s}_n, \mathbf{l}_n)$ denote state and mixture component sequences. In the second step (M-step), we find the new estimate by

$$\hat{W}^{(n)} = \underset{\hat{W}^{(n)}}{\arg\max}\, R_{QB}(\hat{W}^{(n)} \big| W^{(n)}). \tag{9}$$

To solve (8)(9), we may expand the expectation of the cluster membership Ω_c in (8) by discarding the terms, which do not involve $\hat{W}_c^{(n)}$, i.e.

$$\begin{aligned} R_{QB}(\hat{W}_c^{(n)} \big| W_c^{(n)}) \propto -\frac{1}{2}\sum_t \sum_{j,k\in\Omega_c} \gamma_t(j,k)(\mathbf{x}_t^{(n)} - \hat{W}_c^{(n)}\xi_{jk})^T \Sigma_{jk}^{-1} \\ \times (\mathbf{x}_t^{(n)} - \hat{W}_c^{(n)}\xi_{jk}) - \frac{1}{2}\sum_{i=1}^{d}\left[\hat{W}_c^{(n)}(i) - M_c^{(n-1)}(i)\right] \\ \times \Sigma_{ci}^{(n-1)^{-1}}\left[\hat{W}_c^{(n)}(i) - M_c^{(n-1)}(i)\right]^T, \end{aligned} \tag{10}$$

where $\gamma_t(j,k) = \Pr(s_t^{(n)} = j, l_t^{(n)} = k \big| \mathbf{X}_n = \{\mathbf{x}_t^{(n)}\}, W_c^{(n)})$. Assuming that HMM covariance matrices are diagonal, $\Sigma_{jk} = diag(\sigma_{jk1}^2, \cdots, \sigma_{jkd}^2)$, (10) can be further arranged by

$$\begin{aligned} R_{QB}(\hat{W}_c^{(n)} \big| W_c^{(n)}) \propto -\frac{1}{2}\sum_{i=1}^{d}\left[\hat{W}_c^{(n)}(i) - \overline{W}_c(i)\right]\left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2}\xi_{jk}\xi_{jk}^T\right] \\ \times\left[\hat{W}_c^{(n)}(i) - \overline{W}_c(i)\right]^T - \frac{1}{2}\sum_{i=1}^{d}\left[\hat{W}_c^{(n)}(i) - M_c^{(n-1)}(i)\right]\Sigma_{ci}^{(n-1)^{-1}}\left[\hat{W}_c^{(n)}(i) \right. \\ \left. - M_c^{(n-1)}(i)\right]^T \propto -\frac{1}{2}\sum_{i=1}^{d}\left[\hat{W}_c^{(n)}(i) - \hat{M}_c(i)\right]\hat{\Sigma}_{ci}^{-1}\left[\hat{W}_c^{(n)}(i) - \hat{M}_c(i)\right]^T, \end{aligned} \tag{11}$$

with $\mathbf{x}_t^{(n)} = \{x_{ti}^{(n)}\}$ and

$$\overline{W}_c(i) = \left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2} x_{ti}^{(n)}\xi_{jk}^T\right]\left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2}\xi_{jk}\xi_{jk}^T\right]^{-1} \tag{12}$$

We obviously see that the exponential of expectation function $R_{QB}(\hat{W}_c^{(n)} \big| W_c^{(n)})$ is proportional to a matrix variate normal distribution with updated hyperparameters

$$\begin{aligned} \hat{M}_c(i) = \left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2} x_{ti}^{(n)}\xi_{jk}^T + M_c^{(n-1)}(i)\Sigma_{ci}^{(n-1)^{-1}}\right] \\ \times\left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2}\xi_{jk}\xi_{jk}^T + \Sigma_{ci}^{(n-1)^{-1}}\right]^{-1} \end{aligned} \tag{13}$$

$$\hat{\Sigma}_{ci} = \left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2}\xi_{jk}\xi_{jk}^T + \Sigma_{ci}^{(n-1)^{-1}}\right]^{-1}. \tag{14}$$

Therefore, a reproducible prior/posterior density pair is generated. With the property of conjugate density, the maximization (M-step) can be easily performed to obtain QB estimate

$$\hat{W}_{QB}^{(n)} = \hat{M}_c = \{\hat{M}_c(i)\}\,. \tag{15}$$

At the same time, we also construct the mechanism for hyperparameter evolution

$$\varphi^{(n)} = \{\varphi_c^{(n)}\} = \{M_c^{(n)}, \Delta_c^{(n)}\} = \{\hat{M}_c, \hat{\Delta}_c\} = \{\{\hat{M}_c(i)\}, \{\hat{\Sigma}_{ci}\}\}\,. \tag{16}$$

The updated hyperparameters $\varphi^{(n)}$ subsequently serve as new hyperparameters for linear regression adaptation using upcoming data $\mathbf{X}_{n+1}$.

4. SPECIAL REALIZATIONS TO MLLR AND MAPLR

MLLR and MAPLR are both batch adaptation methods, which the regression matrices $W = \{W_c\}$ are estimated via EM algorithm by applying batch adaptation data $\mathbf{X} = \{\mathbf{x}_t\}$. In MLLR, the ML estimate is involved and the resulting expectation function of cluster c is

$$R_{ML}(\hat{W}_c | W_c) \propto -\frac{1}{2}\sum_{i=1}^{d}\left[\hat{W}_c(i) - \overline{W}_c(i)\right] \times \left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2}\xi_{jk}\xi_{jk}^T\right]\left[\hat{W}_c(i) - \overline{W}_c(i)\right]^T. \tag{17}$$

By maximizing (17) with respect to $\hat{W}_c$, we can obtain ML estimate $\hat{W}_{ML,c} = \overline{W}_c = \{\overline{W}_c(i)\}$ with $\overline{W}_c(i)$ shown in (12). This formula is exactly equal to that in MLLR. The notations $G^{(i)}$ and $\mathbf{z}_i$ in MLLR [10] are herein expressed by

$$G^{(i)} = \left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2}\xi_{jk}\xi_{jk}^T\right] \tag{18}$$

$$\mathbf{z}_i = \left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2} x_{ti}^{(n)}\xi_{jk}^T\right]. \tag{19}$$

On the other hand, MAPLR was trying to find MAP estimate of linear regression model by combining the subjective prior $\varphi = \{M_c, \Delta_c\} = \{\{M_c(i)\}, \{\Sigma_{ci}\}\}$, which could be empirically estimated from training data [1][3][5]. Correspondingly, the expectation function $R_{MAP}(\hat{W}_c | W_c)$ has the same form as (11). By maximizing $R_{MAP}(\hat{W}_c | W_c)$ with respect to $\hat{W}_c$, the ith row of MAP linear regression matrix is found to be

$$\hat{W}_{MAP,c}(i) = \left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2} x_{ti}\xi_{jk}^T + M_c(i)\Sigma_{ci}^{-1}\right] \times \left[\sum_t \sum_{j,k\in\Omega_c} \frac{\gamma_t(j,k)}{\sigma_{jki}^2}\xi_{jk}\xi_{jk}^T + \Sigma_{ci}^{-1}\right]^{-1}, \tag{20}$$

which is equivalent to that derived in [5].

However, the proposed QBLR is a sequential adaptation algorithm with special realizations of MLLR and MAPLR. We demonstrate that *the matrix variate normal distribution belongs to the conjugate prior distribution family*. The regression matrices can be optimally estimated at each sequential epoch according to the QB principle. Simultaneously, an attractive scheme for hyperparameter evolution is developed.

5. EXPERIMENTS

5.1 Speech Databases and Baseline System

We conduct a series of connected Chinese digit recognition experiments to examine the goodness of QBLR. Two severely mismatched speech databases were collected [4]. One is the training database consisted of 1000 utterances by 50 males and 50 females. This database was recorded in office environments via four close-talking microphones. We applied this database to train SI HMM's. The second database contained the utterances of five males and five females recorded in two median-class cars: TOYOTA COROLLA 1.8 and YULON SENTRA 1.6. These utterances were collected using a high-quality MD Walkman of type MZ-R55 via a hands-free far talking SONY ECM-717 microphone different from those in training database. Three materials of *standby* condition, *downtown* condition and *freeway* condition with averaged car speeds respectively being 0 km/h, 50 km/h and 90 km/h were recorded. During recording, we kept the engine on, the air-conditioner on, the music off and the windows rolled up. The numbers of testing utterances were 50, 150, 250 and corresponding digits were 324, 964, 1593 for driving conditions of standby, downtown and freeway, respectively. The word error rate (WER) was averaged over ten test speakers. Each speaker had additional five adaptation utterances. All training/testing utterances contained three to eleven random digits. Each Chinese digit was modeled using a seven-state CDHMM. Each HMM state was composed of four mixture components. The feature vector was consisted of 12-order LPC-derived cepstral coefficients, 12-order delta cepstral coefficients, one delta log energy and one delta delta log energy. Our baseline system reports the word error rates of 25.6%, 55% and 62.3% for driving conditions of standby, downtown and freeway, respectively. Only one transformation cluster was adopted.

5.2 Estimation of Initial Hyperparameters

To have a good beginning in online speaker adaptation, the initial hyperparameters play an important role. Herein, the initial hyperparameters $\varphi^{(0)}$ were empirically estimated from SI training data [1]. Let $\chi_1, \cdots, \chi_Q$ denote the training data sets of Q speakers. We first apply these data to calculate the regression matrices $\{\tilde{W}_{c1}\}, \cdots, \{\tilde{W}_{cQ}\}$ corresponding to individual training speakers using the SI HMM's. Then, the initial hyperparameters $\{M_c^{(0)}\}$ and $\{\Sigma_{ci}^{(0)}\}$ are determined by respectively taking the ensemble mean and covariance over the matrices $\{\tilde{W}_{c1}\}, \cdots, \{\tilde{W}_{cQ}\}$, i.e.

$$M_c^{(0)} = \frac{1}{Q}\sum_{q=1}^{Q}\tilde{W}_{cq} \tag{21}$$

$$\Sigma_{ci}^{(0)} = \frac{1}{Q}\sum_{q=1}^{Q}(\tilde{W}_{cq} - M_c^{(0)})^T(\tilde{W}_{cq} - M_c^{(0)})\,. \tag{22}$$

5.3 Comparison of MLLR, MAPLR and QBLR

Two sets of experiments are carried out for evaluation. First, we evaluate the convergence property of online adaptation using QBLR. For each test speaker, we sequentially use five labeled adaptation sentences to perform supervised QBLR adaptation. The adapted HMM parameters are fixed for recognition of his/her test data. As shown in Figure 1, the WER's versus various driving conditions and adaptation data amounts (N) are reported. The baseline system with N=0 is included. We can see that WER's are consistently decreased for various driving conditions when more adaptation data are applied. This indicates the goodness of hyperparameter evolution and QB estimation in QBLR. In case of N=5, WER's have been significantly reduced to 15.3%, 35.7% and 38.3% for standby, downtown and freeway driving conditions.

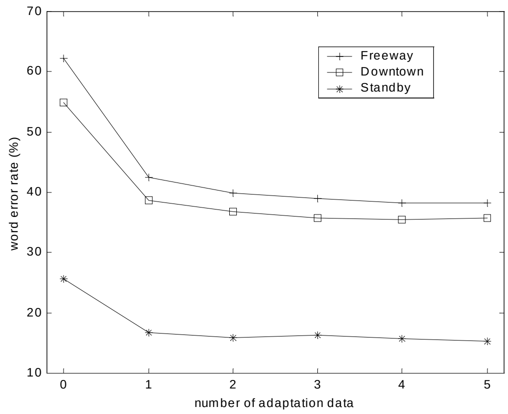

Figure 1: Comparison of WER's (%) using QBLR for various driving conditions and adaptation data amounts.

On the other hand, we also compare the results of baseline system, MLLR, MAPLR and QBLR. Cases of N=1 and N=5 are investigated. As listed in Table 1, the results of MAPLR are better than those of MLLR for case of N=1 and similar to those of MLLR for case of N=5. This phenomenon is observed for various driving conditions. It is because that incorporation of prior statistics is beneficial to parameter estimation when adaptation data is sparse (N=1). But, when more adaptation data (N=5) are involved, the prior information is de-emphasized in estimation of regression matrices. Nevertheless, QBLR still achieves the best results among these methods. Although the improvement is mild, the superiority of QBLR is owing to the adaptation efficiency. Using QBLR with N=5, only one adaptation sentence is needed in each sequential epoch. Efficiency is better than MLLR and MAPLR where all adaptation data are employed in single one epoch.

	Baseline System	MLLR, N=1	MAPLR, N=1	MLLR, N=5	MAPLR, N=5	QBLR, N=5
Standby	25.6	17.1	16.8	16.1	16.7	15.3
Downtown	55	39.8	38.7	36.9	36.7	35.7
Freeway	62.3	43.4	42.4	40	40.3	38.3

Table 1: Comparison of WER's (%) of baseline, MLLR, MAPLR and QBLR.

6. CONCLUSION

This paper has presented a novel QBLR algorithm for online speaker adaptation of HMM parameters. Different from MLLR and MAPLR designed for batch adaptation, QBLR was a sequential adaptation approach to achieve high adaptation efficiency in terms of computation cost and memory requirement. We compared the theoretical models of MLLR, MAPLR and QBLR. In QBLR, the parameter of linear regression model was estimated by maximizing the approximated posterior density of accumulated adaptation data, which was equal to the product of a likelihood of current data and a prior density of regression parameter. Due to QBLR, we knew the underlying theory of sequential adaptation and the mechanisms of how to evolve statistics of linear regression model and employ the statistics to estimate the optimal regression parameter. We carefully selected the prior density of regression matrix to be a matrix variate normal distribution and proved this density belonging to conjugate prior distribution family. Accordingly, we obtained a concise and closed form QB estimate of regression matrix. More importantly, a reproducible prior/posterior density pair was produced to build a meaning mechanism for evolution of regression statistics. In this study, we showed the generalization essence of the derived QBLR formulas with relations to formulas of MLLR and MAPLR. In the experiments of car speech recognition, the convergence property of QBLR is confirmed. Also, the superiority of QBLR over MLLR and MAPLR is demonstrated.

7. REFERENCES

1. Chesta, C., Siohan, O. and C.-H. Lee, "Maximum a posterior linear regression for hidden Markov model adaptation", in *Proc. EUROSPEECH*, vol. 1, pp. 211-214, 1999.
2. Chien, J.-T., "Online hierarchical transformation of hidden Markov models for speech recognition", *IEEE Trans. SAP*, vol. 7, no. 6, pp. 656-667, 1999.
3. Chien, J.-T. and Wang, H.-C., "Telephone speech recognition based on Bayesian adaptation of hidden Markov models", *Speech Communication*, vol. 22, no. 4, pp. 369-384, 1997.
4. Chien, J.-T. and Liao, G.-H., "Transformation-based Bayesian predictive classification for online environmental learning and robust speech recognition", in *Proc. ICSLP*, vol. 4, pp. 544-547, 2000.
5. Chou, W., "Maximum a posterior linear regression with elliptically symmetric matrix variate priors", in *Proc. EUROSPEECH*, vol. 1, pp. 1-4, 1999.
6. Dempster, A. P., Laird, N. M. and Rubin, D. B., "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy. Stat. Soc.*, vol. 39, pp. 1-38, 1977.
7. Giri, N. C., *Multivariate statistical analysis*, Marcel Dekker, Inc., 1996.
8. Gupta, A. K. and Varga, T., *Elliptically Contoured Models in Statistics*, Kluwer Academic Publishers, 1993.
9. Huo, Q. and Lee, C.-H., "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate", *IEEE Trans. SAP*, vol. 5, no. 2, pp. 161-172, 1997.
10. Leggetter, C. J. and Woodland, P. C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.
11. Wang, S. and Zhao, Y., "On-line Bayesian speaker adaptation using tree-structured transformation and robust priors", in *Proc. ICASSP*, vol. 2, pp. 977-980, 2000.