Joint Bilinear Transformation Space Based Maximum a Posteriori Linear Regression Adaptation using Prior with Variance Function

Hwa Jeon Song[1], *Yunkeun Lee*[1], *Hyung Soon Kim*[2]

[1]Spoken Language Processing Team, Electronics and Telecommunications Research Institute, Korea
[2]School of Electrical Engineering, Pusan National University, Korea

hwajeon@etri.re.kr, yklee@etri.re.kr, kimhs@pusan.ac.kr

Abstract

This paper proposes a new joint maximum a posteriori linear regression (MAPLR) adaptation using single prior distribution with a variance function in bilinear transformation space (BITS). There are two indirect adaptation methods based on the linear transformation in BITS and these are tightly coupled by joint MAP-based estimation. The proposed method not only has the scalable parameters but also is based on only one prior distribution, unlike the conventional joint MAP-MAPLR method with two priors. Experimental results, especially for small amount of adaptation data, show the synergy between two indirect BITS-based methods over other methods.

Index Terms: speaker adaptation, bilinear model, maximum a posteriori linear regression (MAPLR), MAP

1. Introduction

Speaker adaptation technique is an effective way for more customized services based on the speech technology such as mobile voice search and voicemail which are recently in the spotlight of mobile services.

Many speaker adaptation methods have been proposed to improve the baseline performance. In fact, an aspect studied in the methods has been deeply related with the adaptation data size. As you know, it is important for the adapted system to keep higher than baseline performance regardless of adaptation data size. For example, on the adapted mean vector $\boldsymbol{\mu}^s$ of a specific speaker s, a simple but powerful expression being free from the constraint of the adaptation data size can be represented as:

$$\boldsymbol{\mu}^s = \alpha\boldsymbol{\mu}^s_{ES} + \beta\boldsymbol{\mu}^s_{MLLR} + \gamma\boldsymbol{\mu}^s_{ML} \tag{1}$$

where $\boldsymbol{\mu}^s_{ES}$, $\boldsymbol{\mu}^s_{MLLR}$ and $\boldsymbol{\mu}^s_{ML}$ denote eigenspace-based maximum likelihood linear regression(ES-MLLR) [1], MLLR [2] and ML estimates, respectively; α, β and γ are automatically adjusted according to the adaptation data size; $\alpha + \beta + \gamma = 1$. Actually, as the systematic scheme to control α, β and γ in (1), the maximum a posteriori (MAP) [3] criterion has been used in many approaches [1, 4, 5, 6]. MAPLR approach [1, 5] including *joint* MAP-MAPLR scheme [6] is a typical MAP-based approach as the Bayesian counterpart of ES-MLLR and MLLR.

This paper, for an effective realization of (1), proposes a new *joint* bilinear transformation space-based MAPLR (BIT-MAPLR) framework using the prior distribution with a variance function in bilinear transformation space (BITS) under the conventional MAP criterion. It is based on the recently proposed adaptation methods [7, 8] using the bilinear model (BM) [9] concept, separating two independent variations in a set of observations. Moreover, the proposed scheme can be viewed as a generalized framework encompassing MLLR, ES-MLLR, MAP, MAPLR and joint MAP-MAPLR.

Section 2 describes the BIT-MLLR family; Section 3 and 4 respectively describe the joint BIT-MAPLR with two priors and the proposed method with single prior; After reporting the experimental results in Section 5, Section 6 has a conclusion.

2. Bilinear Transformation Space-based Speaker Adaptation Framework

2.1. Decomposition of Transformation Matrices in BITS

Let the speaker independent (SI) hidden Markov model (HMM) be built by using training database of S speakers, resulting in N states with K mixtures per state; $\boldsymbol{\mu}_{nk}$ denotes D-dimensional mean vector of Gaussian mixture k in state n but $\boldsymbol{\mu}_c$ with *content* c is used instead of $\boldsymbol{\mu}_{nk}$ with state/mixture nk for notational simplicity, where $1 \leq c \leq C\ (= N \times K)$. Assume that the relation between $\boldsymbol{\mu}^s_c$ of speaker s and $\boldsymbol{\mu}_c$ is $\boldsymbol{\mu}^s_c = \mathbf{A}^s\boldsymbol{\mu}_c + \mathbf{b}^s = \mathbf{W}^s\boldsymbol{\xi}_c$ where $\mathbf{W}^s = [\mathbf{b}^s\ \mathbf{a}^s_1 \cdots \mathbf{a}^s_D] \in \mathbb{R}^{D\times(D+1)}$ and $\mathbf{a}^s_d$ denotes the dth column vector of $\mathbf{A}^s$; $\boldsymbol{\xi}_c = [1\ \boldsymbol{\mu}^T_c]^T$.

From these S speaker-dependent (SD) transformation matrices, two independent factors reflecting the variable and invariant components over all speakers in BITS can be obtained by using two types of stacked matrices as

$$\overline{\mathbf{W}} = \left[\mathbf{W}^{1^T} \cdots \mathbf{W}^{S^T}\right]^T, \quad \overline{\mathbf{W}}^{VT} = \left[\mathbf{w}^1 \cdots \mathbf{w}^S\right] \tag{2}$$

where ^{VT} denotes the vector transpose of any matrix [9]; $\overline{\mathbf{W}} \in \mathbb{R}^{SD\times(D+1)}$ and $\overline{\mathbf{W}}^{VT} \in \mathbb{R}^{(D+1)D\times S}$; $\mathbf{w}^s = [\mathbf{b}^{s^T}\ \mathbf{a}^{s^T}_1 \cdots \mathbf{a}^{s^T}_D]^T$ is a meta-vector built by concatenating the $D{+}1$ ordered column vectors of $\mathbf{W}^s$. That is, the two matrices in (2) are decomposed under the symmetric BM building procedure in [9] as follows:

$$\overline{\mathbf{W}} = \left[\mathbf{M}^{VT}\mathbf{S}\right]^{VT}\mathbf{Q} = \left[[\mathbf{MQ}]^{VT}\,\mathbf{S}\right]^{VT} \tag{3}$$

where $\mathbf{S} \in \mathbb{R}^{I\times S}$, $\mathbf{Q} \in \mathbb{R}^{J\times(D+1)}$, $\mathbf{M} \in \mathbb{R}^{(ID)\times J}$, $I \leq S$ and $J \leq D+1$; $\mathbf{S} = [\mathbf{s}^1 \cdots \mathbf{s}^S]$ is the style basis matrix; $\mathbf{Q}$ is the orthonormal basis of eigenvector to reduce the dimensionality of the canonical model such as SI HMM; Note that $\mathbf{s}^s$ in $\mathbf{S}$, which is called a *style factor*, reflects the unique characteristic of speaker s, while $\mathbf{Q}$ is invariant across the speakers; $\mathbf{M}$ is the bilinear mapping matrix to control two factors independently. From (3), the transformation matrix(TM) of each speaker s can be represented $\mathbf{W}^s = [\mathbf{M}^{VT}\mathbf{s}^s]^{VT}\mathbf{Q}$ or $\mathbf{w}^s = [\mathbf{MQ}]^{VT}\mathbf{s}^s$ in BITS. As such, $\boldsymbol{\mu}^s_c$ also can be written as

$$\boldsymbol{\mu}^s_c = \left[\mathbf{M}^{VT}\mathbf{s}^s\right]^{VT}\mathbf{Q}\boldsymbol{\xi}_c = [\mathbf{MQ}\boldsymbol{\xi}_c]^{VT}\,\mathbf{s}^s. \tag{4}$$

Given the adaptation data of a new speaker, with a style $\tilde{s}$, composed of T observations, $\mathbf{O} = \{\mathbf{o}_t, t = 1, \cdots, T\}$, the adapted model can be estimated based on ML estimation. Depending on a kind of style factor, there are two types of speaker adaptation schemes. For the alternative approach, refer [8].

2.2. BIT-MLLR by Transform

BIT-MLLR *by transform* (BIT-MLLR$_T$) is briefly described. In BIT-MLLR$_T$, the new speaker's model from (4) is written by

$$\boldsymbol{\mu}^{\tilde{s}}_{c,T} = \Big[\mathbf{M}^{VT}\mathbf{s}^{\tilde{s}}_T\Big]^{VT}\mathbf{Q}\boldsymbol{\xi}_c = \mathbf{X}^{\tilde{s}}_T\mathbf{Q}\boldsymbol{\xi}_c \tag{5}$$

where $\mathbf{X}^{\tilde{s}}_T$=$[\mathbf{M}^{VT}\mathbf{s}^{\tilde{s}}_T]^{VT} \in \mathbb{R}^{D\times J}$; '$_T$' denotes the transform-based method like as MLLR. In BIT-MLLR$_T$, $\mathbf{X}^{\tilde{s}}_T$ itself instead of $\mathbf{s}^{\tilde{s}}_T$ should be estimated for more accurate specific-speaker model. Since $\mathbf{Q}$ is fixed across the speakers, $\mathbf{X}^{\tilde{s}}_T$ can be found as $\hat{\mathbf{X}}^{\tilde{s}}_T$=argmax $_{\mathbf{X}^{\tilde{s}}_T} \log p\big(\mathbf{O} \mid \mathbf{X}^{\tilde{s}}_T\big)$. This is equivalent to the estimation procedure in MLLR [2] except for the dimension reduction from $\boldsymbol{\xi}_c$ to $\mathbf{Q}\boldsymbol{\xi}_c$. Moreover, when J=D+1, BIT-MLLR$_T$ becomes equivalent to MLLR since no dimension reduction or no information loss by linear transformation based on $\mathbf{Q}$ from $\boldsymbol{\xi}_c$ is occurred.

2.3. BIT-MLLR by Projection

Now, BIT-MLLR *by projection* (BIT-MLLR$_P$) is discussed. The adapted model in BIT-MLLR$_P$ from (4) is given by

$$\boldsymbol{\mu}^{\tilde{s}}_{c,P} = [\mathbf{M}\mathbf{Q}\boldsymbol{\xi}_c]^{VT}\,\mathbf{s}^{\tilde{s}}_P \tag{6}$$

where $\mathbf{s}^{\tilde{s}}_P$ is the style factor of the new speaker to be estimated and '$_P$' for projection-based method in $\hat{\boldsymbol{\mu}}^{\tilde{s}}_{c,P}$ and $\mathbf{s}^{\tilde{s}}_P$ is used to distinguish from BIT-MLLR$_T$. Therefore the style factor $\mathbf{s}^{\tilde{s}}_P$ can be found as $\hat{\mathbf{s}}^{\tilde{s}}_P$=argmax $_{\mathbf{s}^{\tilde{s}}_P} \log p\big(\mathbf{O} \mid \mathbf{s}^{\tilde{s}}_P\big)$. This is the same estimation formula in ES-MLLR. Like BIT-MLLR$_T$, BIT-MLLR$_P$ becomes equivalent to ES-MLLR when J=D+1. Moreover, note that (6) can be rewritten as $\boldsymbol{\mu}^{\tilde{s}}_{c,P}$=$\mathbf{X}^{\tilde{s}}_P\mathbf{Q}\boldsymbol{\xi}_c$ where $\mathbf{X}^{\tilde{s}}_P$=$[\mathbf{M}^{VT}\mathbf{s}^{\tilde{s}}_P]^{VT}$.

3. Joint BIT-MAPLR

This section explains the joint BIT-MAPLR whereby $\mathbf{X}^{\tilde{s}}$ and $\mathbf{Q}$ are jointly estimated. In fact, each component of $\mathbf{Q}$ is variable owing to the property of BM with two interchangeable variations. Firstly, for more detailed transform on the content c of the new speaker, $\mathbf{Q}_c$ instead of $\mathbf{Q}$ in (5) is used as follows:

$$\boldsymbol{\mu}^{\tilde{s}}_c = \mathbf{X}^{\tilde{s}}\mathbf{Q}_c\boldsymbol{\xi}_c = \mathbf{X}^{\tilde{s}}\mathbf{z}_c \tag{7}$$

where $\mathbf{z}_c$=$\mathbf{Q}_c\boldsymbol{\xi}_c$. Then, let the parameter set is defined as $\Theta = \{\lambda, \eta\}$ for $\mathbf{X}^{\tilde{s}}$ and $\mathbf{Q}_c$. Lastly, under MAP-based criterion, the auxiliary function with respect to the re-estimated set $\hat{\Theta}$ in EM algorithm [10] is defined as

$$R(\hat{\Theta} \mid \Theta) = E\big[\log\{P(\mathbf{O}, \mathcal{C} \mid \hat{\lambda}, \hat{\eta})P(\hat{\lambda}, \hat{\eta})\} \mid \mathbf{O}, \lambda, \eta\big] \tag{8}$$

where $E[\cdot]$ denotes the statistical expectation; $\mathcal{C}$ represents the content sequence (i.e., joint state and mixture sequences); $P(\hat{\lambda}, \hat{\eta})$ is a joint prior probability density function (pdf) of λ and η. Since style and content factors are independent of each other in BITS, $P(\hat{\lambda}, \hat{\eta}) = P(\hat{\lambda})P(\hat{\eta})$. Moreover, though there are not obvious conjugate prior pdfs in MAPLR, the matrix variate normal density is generally used as the prior pdf [5, 6].

If the information on the style of the new speaker could be incorporated into the prior model in advance, it is quite effective to improve the performance. Actually, since we can obtain the good quality model using BIT-MLLR$_P$ even for a very limited adaptation data, we assume the prior pdfs for $\mathbf{X}^{\tilde{s}}$ and $\mathbf{Q}_c$ as follows: The prior pdf for $\mathbf{X}^{\tilde{s}}$ is $P(\hat{\mathbf{X}}^{\tilde{s}}) = \mathcal{N}_M(\hat{\mathbf{X}}^{\tilde{s}}; \mathbf{X}^{\tilde{s}}_P, \boldsymbol{\Omega}_X, \boldsymbol{\tau}_s)$ where $\mathcal{N}_M(;)$ denotes the matrix variate normal density [6]; $\mathbf{X}^{\tilde{s}}_P$ is the mean matrix; both $\boldsymbol{\Omega}_X \in \mathbb{R}^{J\times J}$ and $\boldsymbol{\tau}_s \in \mathbb{R}^{D\times D}$ are the covariance matrices; The prior pdf for $\mathbf{Q}_c$ is $P(\hat{\mathbf{Q}}_c) = \mathcal{N}_M(\hat{\mathbf{Q}}_c; \mathbf{Q}, \boldsymbol{\Omega}_Q, \boldsymbol{\tau}_c)$ with $\mathbf{Q}$ as the mean matrix, where $\boldsymbol{\Omega}_Q \in \mathbb{R}^{(D+1)\times(D+1)}$ and $\boldsymbol{\tau}_c \in \mathbb{R}^{J\times J}$ as the covariance matrices; The hyperparameters $\boldsymbol{\Omega}_X$, $\boldsymbol{\tau}_s$, $\boldsymbol{\Omega}_Q$ and $\boldsymbol{\tau}_c$ are assumed to have the non-informative priors because they are not easy to specify. In fact, if a latent variable model could be adopted on $\mathbf{X}^{\tilde{s}}$ and $\mathbf{Q}_c$, it becomes more analytical. Also, notice that two above prior pdfs are *homogeneous* each other contrary to *heterogeneous* ones in [6]. After ignoring all terms independent of $\hat{\mathbf{X}}^{\tilde{s}}$ and $\hat{\mathbf{Q}}_c$, (8) can be rearranged as

$$\begin{aligned} R(\hat{\Theta} \mid \Theta) = \sum_{t=1}^{T}\sum_{c=1}^{C}\gamma_c(t)\Big[-\frac{1}{2}\big(\mathbf{o}_t - \hat{\boldsymbol{\mu}}^{\tilde{s}}_c\big)^T\boldsymbol{\Sigma}_c^{-1}\big(\mathbf{o}_t - \hat{\boldsymbol{\mu}}^{\tilde{s}}_c\big)\Big] \\ + f(\hat{\mathbf{X}}^{\tilde{s}}; \mathbf{X}^{\tilde{s}}_P, \boldsymbol{\tau}_s, \boldsymbol{\Omega}_X) + f(\hat{\mathbf{Q}}_c; \mathbf{Q}, \boldsymbol{\tau}_c, \boldsymbol{\Omega}_Q) \end{aligned} \tag{9}$$

where $f(\mathbf{A}; \mathbf{B}, \mathbf{C}, \mathbf{D}) = -1/2\, tr(\mathbf{A}-\mathbf{B})^T\mathbf{C}^{-1}(\mathbf{A}-\mathbf{B})\mathbf{D}^{-1}$ and $tr(\cdot)$ is the trace operator; for convenience, $\boldsymbol{\Omega}_X$ and $\boldsymbol{\Omega}_Q$ are chosen as the identity matrices.

Though there is no way to maximize the likelihood with respect to $\hat{\mathbf{X}}^{\tilde{s}}$ and $\hat{\mathbf{Q}}_c$ simultaneously, the parameters can be separately estimated by the iterative MAP principle [6]. For more details, refer [6]. Firstly, let $\hat{\lambda}$ be estimated with fixed η. That is, for estimating $\hat{\mathbf{X}}^{\tilde{s}}$, we set as $\hat{\boldsymbol{\mu}}^{\tilde{s}}_c = \hat{\mathbf{X}}^{\tilde{s}}\mathbf{Q}_c\boldsymbol{\xi}_c$ in (9) and then $\partial R(\hat{\lambda}, \eta \mid \lambda, \eta)/\partial\hat{\mathbf{X}}^{\tilde{s}} = \mathbf{0}$. Through some arrangements, the following is obtained:

$$\sum_{c=1}^{C}\mathbf{V}_c\hat{\mathbf{X}}^{\tilde{s}}\mathbf{Z}_c + \boldsymbol{\tau}_s^{-1}\hat{\mathbf{X}}^{\tilde{s}} = \sum_{c=1}^{C}\mathbf{k}_c\mathbf{z}_c^T + \boldsymbol{\tau}_s^{-1}\mathbf{X}^{\tilde{s}}_P \tag{10}$$

where $\mathbf{k}_c = \sum_{t=1}^{T}\gamma_c(t)\boldsymbol{\Sigma}_c^{-1}\mathbf{o}_t$; $\mathbf{V}_c = \sum_{t=1}^{T}\gamma_c(t)\boldsymbol{\Sigma}_c^{-1}$; $\mathbf{Z}_c = \mathbf{z}_c\mathbf{z}_c^T$; assume that both $\boldsymbol{\tau}_s$ and $\boldsymbol{\Sigma}_c$ are diagonal. This is basically equivalent to the conventional MAPLR [5] and we call this method BIT-MAPLR.

Now, $\hat{\lambda}$ in the above procedure is fixed to estimate $\hat{\eta}$. That is, $\hat{\mathbf{Q}}_c$ can be obtained by setting as $\hat{\boldsymbol{\mu}}_c = \hat{\mathbf{X}}^{\tilde{s}}\hat{\mathbf{Q}}_c\boldsymbol{\xi}_c$ in (9). After setting $\boldsymbol{\Psi}_c = \hat{\mathbf{X}}^{\tilde{s}^T}\mathbf{V}_c\hat{\mathbf{X}}^{\tilde{s}}$ and $\boldsymbol{\kappa}_c = \hat{\mathbf{X}}^{\tilde{s}^T}\mathbf{k}_c$, then $\partial R(\hat{\lambda}, \hat{\eta} \mid \hat{\lambda}, \eta)/\partial\hat{\mathbf{Q}}_c = \mathbf{0}$. Hence, the following is obtained:

$$\boldsymbol{\Psi}_c\hat{\mathbf{Q}}_c\boldsymbol{\xi}_c\boldsymbol{\xi}_c^T + \boldsymbol{\tau}_c^{-1}\hat{\mathbf{Q}}_c = \boldsymbol{\kappa}_c\boldsymbol{\xi}_c^T + \boldsymbol{\tau}_c^{-1}\mathbf{Q}. \tag{11}$$

Notice that (11) is basically equivalent to (10) by homogeneous property between $\hat{\mathbf{X}}^{\tilde{s}}$ and $\hat{\mathbf{Q}}_c$. However, it seems to be no way to solve for $\hat{\mathbf{Q}}_c$ in (11) because $\boldsymbol{\Psi}_c$ is not necessarily diagonal unlike $\mathbf{V}_c$ in (10). Instead of that, $\hat{\mathbf{Q}}_c$ can be indirectly obtained as follows: After setting $\mathbf{D}_c = \boldsymbol{\xi}_c\boldsymbol{\xi}_c^T$, let both sides of (11) on the right be multiplied by $\mathbf{D}_c$ as

$$\boldsymbol{\Psi}_c\hat{\mathbf{Q}}_c\mathbf{D}_c\mathbf{D}_c + \boldsymbol{\tau}_c^{-1}\hat{\mathbf{Q}}_c\mathbf{D}_c = \boldsymbol{\kappa}_c\boldsymbol{\xi}_c^T\mathbf{D}_c + \boldsymbol{\tau}_c^{-1}\mathbf{Q}\mathbf{D}_c. \tag{12}$$

Notice that $\mathbf{D}_c\mathbf{D}_c = \sigma_c\mathbf{D}_c$ where $\sigma_c = \boldsymbol{\xi}_c^T\boldsymbol{\xi}_c$. Hence,

$$\hat{\mathbf{Q}}_c\mathbf{D}_c = \Big[\sigma_c\boldsymbol{\Psi}_c + \boldsymbol{\tau}_c^{-1}\Big]^{-1}\Big[\boldsymbol{\kappa}_c\boldsymbol{\xi}_c^T + \boldsymbol{\tau}_c^{-1}\mathbf{Q}\Big]\mathbf{D}_c. \tag{13}$$

Though $\mathbf{D}_c$ is necessarily not invertible, $\hat{\mathbf{Q}}_c = [\sigma_c\boldsymbol{\Psi}_c + \boldsymbol{\tau}_c^{-1}]^{-1}[\boldsymbol{\kappa}_c\boldsymbol{\xi}_c^T + \boldsymbol{\tau}_c^{-1}\mathbf{Q}]$ is one of the solutions for equation equality. This approach based on (10) and (13) is called *joint* BIT-MAPLR ($\jmath$BIT-MAPLR).

4. Joint BIT-MAPLR using Single Prior with a Variance Function

Two BIT-MLLRs share the contents as the common canonical model. To utilize this advantage, this paper proposes a new MAPLR approach in BITS based on the conventional MAP approach, instead of MAPLR approach in the previous section. Unlike MAPLR, there are the various prior pdfs that should satisfy the conjugate family of the complete-data density in MAP. For a good quality prior, assume that the prior pdf of $\hat{\boldsymbol{\mu}}_c^{\tilde{s}}$ is $P(\hat{\boldsymbol{\mu}}_c^{\tilde{s}}) = \mathcal{N}(\hat{\boldsymbol{\mu}}_c^{\tilde{s}}; \boldsymbol{\mu}_{c,P}^{\tilde{s}}, \nu(\theta)\mathbf{I} \mid \lambda, \eta)$, where $\mathcal{N}(;)$ is the multivariate normal density; $\boldsymbol{\mu}_{c,P}^{\tilde{s}} = [\mathbf{MQ}\boldsymbol{\xi}_c]^{VT}\,\mathbf{s}_P^{\tilde{s}}$ is the estimate by BIT-MLLR$_P$ as a mean vector; $\nu(\theta)\mathbf{I}$ is the covariance matrix with a variance function $\nu(\theta)$ which depends on the kind of the estimated parameters (λ, η) and $\mathbf{I}$ denotes the $D \times D$ identity matrix in this paper. Though the prior pdf $P(\hat{\boldsymbol{\mu}}_c^{\tilde{s}})$ can be regarded as the conjugate family, the probabilistic models of $\hat{\mathbf{X}}^{\tilde{s}}$ and $\hat{\mathbf{Q}}_c$ are not obviously defined like [4] with a latent variable model. However, for simplicity, we consider only the covariance term $\nu(\theta)$ depending on the associated variation. Moreover, the MAP criterion in (14) is still valid with improper priors since the only constraint is that the prior pdf should be a non-negative function [5]. Under these assumptions, the new MAP criterion for joint BIT-MAPLR can be defined as

$$\begin{aligned} R_\nu(\hat{\Theta} \mid \Theta) = & \sum_{t=1}^{T}\sum_{c=1}^{C} \gamma_c(t)\Big[-\frac{1}{2}\big(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_c^{\tilde{s}}\big)^T \boldsymbol{\Sigma}_c^{-1}\big(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_c^{\tilde{s}}\big)\Big] \\ & + \sum_{c=1}^{C} g\big(\hat{\boldsymbol{\mu}}_c^{\tilde{s}}; \boldsymbol{\mu}_{c,P}^{\tilde{s}}, \nu(\theta)\mathbf{I} \mid \lambda, \eta\big) \qquad (14) \end{aligned}$$

where $g(\boldsymbol{a}; \boldsymbol{b}, \mathbf{C}) = -1/2\, tr(\boldsymbol{a} - \boldsymbol{b})^T \mathbf{C}^{-1}(\boldsymbol{a} - \boldsymbol{b})$. In fact, differentiating $R_\nu(\cdot)$ with respect to $\hat{\boldsymbol{\mu}}_c^{\tilde{s}}$ as $\partial R_\nu(\cdot)/\partial\hat{\boldsymbol{\mu}}_c^{\tilde{s}} = \mathbf{0}$ is equivalent to MAP adaptation. In addition, if a probabilistic model like a latent variable model on $\boldsymbol{\mu}_{c,P}^{\tilde{s}}$ can be introduced in $g(\cdot)$ of (14), this is basically equivalent to MAP-based method developed in [4] except for some differences between them, including the corresponding training procedures for $[\mathbf{MQ}\boldsymbol{\xi}_c]^{VT}$ in $\boldsymbol{\mu}_{c,P}^{\tilde{s}}$ and basis vectors in [4]. However, because we have to set up more complicate probabilistic model for two variations, it is left for future work.

For simplicity, we chose an intuitive variance function as $\nu(\theta) = \sigma_\lambda\delta(\lambda) + \sigma_\eta\delta(\eta)$, where both σ_λ and σ_η are any constant values and $\delta(\cdot)$ is the Kronecker delta. Firstly, under the iterative MAP principle, $\hat{\mathbf{X}}^{\tilde{s}}$ can be obtained by setting as $\hat{\boldsymbol{\mu}}_c^{\tilde{s}} = \hat{\mathbf{X}}^{\tilde{s}}\mathbf{Q}_c\boldsymbol{\xi}_c = \hat{\mathbf{X}}^{\tilde{s}}\mathbf{z}_c$, $\boldsymbol{\mu}_{c,P}^{\tilde{s}} = [\mathbf{MQ}\boldsymbol{\xi}_c]^{VT}\mathbf{s}_P^{\tilde{s}}$ and $\nu(\lambda) = \sigma_\lambda$ for $\hat{\lambda}$ in (14). From $\partial R_\nu(\cdot)/\partial\hat{\mathbf{X}}^{\tilde{s}} = \mathbf{0}$, the estimation formula is given by

$$\sum_{c=1}^{C}\Big[\mathbf{V}_c\hat{\mathbf{X}}^{\tilde{s}}\mathbf{Z}_c + \frac{1}{\sigma_\lambda}\hat{\mathbf{X}}^{\tilde{s}}\mathbf{Z}_c\Big] = \sum_{c=1}^{C}\Big[\mathbf{k}_c\mathbf{z}_c^T + \frac{1}{\sigma_\lambda}\mathbf{X}_P^{\tilde{s}}\mathbf{Z}_c\Big]. \qquad (15)$$

This method based on (15) is called BIT-MAPLR using the single prior with a variance function (BIT-MAPLR$_\nu$). Unlike (10) in BIT-MAPLR, $g(\cdot)$ in (14) contributes to find $\hat{\mathbf{X}}^{\tilde{s}}$ in (15) by minimizing the summation of differences between the estimated mean vector and the prior mean vector over all content vectors, because all contents share the same specific-style TM. Actually, the second term in (14) is more reasonable than the corresponding one in (9) in that $\hat{\mathbf{X}}^{\tilde{s}}$ is estimated by maximizing the similarity between two style-specific matrices over all contents (i.e., $\sum_c \|\hat{\boldsymbol{\mu}}_c^{\tilde{s}} - \boldsymbol{\mu}_{c,P}^{\tilde{s}}\|^2 = \sum_c \|(\hat{\mathbf{X}}^{\tilde{s}}\mathbf{Q}_c - \mathbf{X}_P^{\tilde{s}}\mathbf{Q})\boldsymbol{\xi}_c\|^2$ instead of $\|\hat{\mathbf{X}}^{\tilde{s}} - \mathbf{X}_P^{\tilde{s}}\|^2$ in BIT-MAPLR).

Now it is time to estimate $\hat{\mathbf{Q}}_c$ by the same procedure in the previous section. However, for convenience, $\hat{\mathbf{Q}}_c$ can be indirectly estimated by the equivalent parameter $\hat{\mathbf{z}}_c$ which is the linear transformation of $\boldsymbol{\xi}_c$ ($\hat{\mathbf{z}}_c = \hat{\mathbf{Q}}_c\boldsymbol{\xi}_c$). Hence, after differentiating as $\partial R_\nu(\cdot)/\partial\hat{\mathbf{z}}_c = \mathbf{0}$ with $\hat{\boldsymbol{\mu}}_c^{\tilde{s}} = \hat{\mathbf{X}}^{\tilde{s}}\hat{\mathbf{z}}_c$, $\boldsymbol{\mu}_{c,P}^{\tilde{s}} = \mathbf{X}_P^{\tilde{s}}\mathbf{Q}\boldsymbol{\xi}_c$ and $\nu(\eta) = \sigma_\eta$ for $\hat{\eta}$, $\hat{\mathbf{z}}_c$ is given as

$$\hat{\mathbf{z}}_c = \Big[\boldsymbol{\Psi}_c + \frac{1}{\sigma_\eta}\mathbf{E}\Big]^{-1}\Big[\hat{\mathbf{X}}^{\tilde{s}^T}\mathbf{k}_c + \frac{1}{\sigma_\eta}\hat{\mathbf{X}}^{\tilde{s}^T}\mathbf{X}_P^{\tilde{s}}\mathbf{Q}\boldsymbol{\xi}_c\Big] \qquad (16)$$

where $\mathbf{E} = \hat{\mathbf{X}}^{\tilde{s}^T}\hat{\mathbf{X}}^{\tilde{s}}$. This approach based on both (15) and (16) is called jBIT-MAPLR$_\nu$.

Now, we have some discussions on our target in (1) with both $\hat{\mathbf{X}}^{\tilde{s}}$ and $\hat{\mathbf{z}}_c$. Firstly, in (10) and (15), the estimate $\hat{\mathbf{X}}^{\tilde{s}}$ can be approximated to $\hat{\mathbf{X}}^{\tilde{s}} \cong \big[\alpha\mathbf{X}_P^{\tilde{s}} + (1-\alpha)\mathbf{X}_T^{\tilde{s}}\big]$ where $\alpha = \tau_\alpha/(\tau_\alpha + \sum_t \gamma_c(t))$ and τ_α is a constant. Secondly, in (13) and (16), $\hat{\mathbf{z}}_c$ is also approximately rewritten as $\hat{\mathbf{z}}_c \cong \beta\mathbf{z}_c + (1-\beta)\mathbf{z}_{c,ML}$, where $\beta = \tau_\beta/(\tau_\beta + \sum_t \gamma_c(t))$ (τ_β is also a constant); $\mathbf{z}_{c,ML} = \hat{\mathbf{X}}^{\tilde{s}^T}\hat{\boldsymbol{\mu}}_{c,ML}^{\tilde{s}}$ and $\hat{\boldsymbol{\mu}}_{c,ML}^{\tilde{s}} = \sum_t \gamma_c(t)\mathbf{o}_t/\sum_t \gamma_c(t)$ denotes the ML estimate. That is, $\hat{\mathbf{z}}_c$ is a MAP estimate in BITS with the reduced dimension by a linear transformation of ML estimate. From the above expressions, $\hat{\boldsymbol{\mu}}_c^{\tilde{s}}$ in (7) can be rearranged as

$$\begin{aligned} \hat{\boldsymbol{\mu}}_c^{\tilde{s}} &= \hat{\mathbf{X}}^{\tilde{s}}\hat{\mathbf{z}}_c \cong \big[\alpha\mathbf{X}_P^{\tilde{s}} + (1-\alpha)\mathbf{X}_T^{\tilde{s}}\big]\big[\beta\mathbf{z}_c + (1-\beta)\mathbf{z}_{c,ML}\big] \\ &= \alpha\beta\mathbf{X}_P^{\tilde{s}}\mathbf{z}_c + (1-\alpha)\beta\mathbf{X}_T^{\tilde{s}}\mathbf{z}_c + (1-\beta)\hat{\mathbf{X}}^{\tilde{s}}\mathbf{z}_{c,ML} \qquad (17) \\ &\cong \kappa_1\boldsymbol{\mu}_{c,P}^{\tilde{s}} + \kappa_2\boldsymbol{\mu}_{c,T}^{\tilde{s}} + \kappa_3\boldsymbol{\mu}_{c,ML}^{\tilde{s}} \qquad (18) \end{aligned}$$

where $\kappa_1 = \alpha\beta$, $\kappa_2 = (1-\alpha)\beta$ and $\kappa_3 = (1-\beta)$; $\kappa_1 + \kappa_2 + \kappa_3 = 1$; To satisfy the equality between (17) and (18) requires $\hat{\mathbf{X}}^{\tilde{s}}\hat{\mathbf{X}}^{\tilde{s}^T} = \mathbf{I}$ in $\boldsymbol{\mu}_{c,ML}^{\tilde{s}}$ term. Hence, $\hat{\boldsymbol{\mu}}_c^{\tilde{s}}$ is corresponded to the dominant one of BIT-MLLR$_P$, BIT-MLLR$_T$ and ML estimates depending on α and β (or amount of adaptation data), which meets the goal in (1).

5. Experiments and Results

To evaluate the performance of the proposed method, we conducted the vocabulary-independent isolated word recognition experiments. We used the Korean phonetically optimized words database (POW DB) provided by ETRI, Korea for training which consists of the speech of about 14 hours collected from 80 speakers (40 males and 40 females). The raw speech was parameterized as the 36-dimensional MFCC vectors (12 MFCCs, their delta and delta-delta coefficients) at every 10 ms over Hamming window of 20 ms. From this, SI HMMs consisted of 3380 tied-states with four Gaussian mixtures per state was constructed and triphone model was used. Then, we obtained 80 SD TMs by MLLR adaptation on each speaker from the SI HMM and each TM was normalized by subtracting the average matrix. Finally, $\mathbf{S}$, $\mathbf{Q}$ and $\mathbf{M}$ were obtained by the symmetric BM building procedure [9] with two stacked matrices in (2). Here, S=80, D=36 and C=13520 (= 3380 × 4).

For the adaptation and the evaluation, we used the Korean phonetically balanced words database (PBW DB) released by SITEC, Korea which was collected from 70 speakers (38 males and 32 females) in a different environment from the POW DB. All of the speakers in PBW DB have uttered the same lexicon composed of 452 words (the average duration per word is about 1 sec). We used 1 to 50 words for the adaptation in a supervised mode, and the remaining 400 words except for the last 2 words are used for the evaluation on each speaker(i.e., 400 isolated words recognition task for total 28000 utterances $= 400 \times 70$). Here, the word accuracy of the baseline system is 96.05%.

In Table 1, the performance of BIT-MLLR family is shown. Notice that BIT-MLLR$_P(I/37)$ and BIT-MLLR$_T(37)$ lead to the recognition performance identical to ES-MLLR(I) and MLLR, respectively. BIT-MLLR$_P$ shows the good performance improvement for small adaptation data while BIT-MLLR$_T$ provides the good improvement over BIT-MLLR$_P$ according as data size is enough large.

Table 1: Word accuracy (%) of BIT-MLLR family

Adaptation Framework	Number of adaptation words						
	1	5	10	20	30	40	50
BIT-MLLR$_P(1/37)$	**96.15**	96.18	96.19	96.18	96.18	96.18	96.17
BIT-MLLR$_T(1)$	94.88	**96.29**	**96.44**	**96.53**	**96.61**	**96.62**	**96.62**
BIT-MLLR$_P(20/37)$	***97.34***	**97.78**	97.80	97.85	97.85	97.83	97.82
BIT-MLLR$_T(20)$	39.78	97.60	**98.44**	**98.64**	**98.65**	**98.61**	**98.64**
BIT-MLLR$_P(30/37)$	***97.39***	**97.92**	97.97	98.04	98.01	98.03	98.03
BIT-MLLR$_T(30)$	-	96.72	***98.45***	**98.67**	**98.73**	**98.71**	**98.72**
BIT-MLLR$_P(37/37)$	**97.33**	***97.98***	98.04	98.07	98.08	98.09	98.07
BIT-MLLR$_T(37)$	-	94.58	**98.28**	***98.70***	***98.74***	***98.77***	***98.78***

In Table 2, we verify the effectiveness of the BIT-MAPLR$_\nu$ family in comparison to BIT-MAPLR family. In the table, 'MAP$_\text{X}$' denotes the standard MAP taking the model obtained from a specific method 'X' as prior. From MAP experiments with various priors, we can observe the importance of the quality of the prior model in MAP. Especially, MAP$_\text{SI}$ has the slow performance degradation because of adapting only a fraction of the poor quality of prior.

All BIT-MAPLR family still works even when adaptation data is very limited like BIT-MLLR$_P$ and ES-MLLR, while MLLR and BIT-MLLR$_T$ with large J have the severe problem in that case. We can also observe that BIT-MAPLR$_\nu$ family leads to consistent improvement over BIT-MAPLR family for small amount of data (1 to 5 adaptation words), irrespective of the number of J because of the smoothing procedure over all contents. Especially, not only does BIT-MAPLR$_\nu$ with $J \geq 30$ lead to recognition performance that is nearly comparable to or better than MLLR (or BIT-MLLR(37)) for more than 20 adaptation words, but it provides considerable improvement for small data size (1 to 10 adaptation words). Moreover, when amount of adaptation data keeps increasing, the BIT-MAPLR$_\nu$ and jBIT-MAPLR$_\nu$ estimates (including BIT-MAPLR family) converge asymptotically to the BIT-MLLR$_T$ estimate and the SD model estimate, respectively.

However, as pointed out in [6], we also observed some numerical instabilities by the inverse operation of the corresponding terms in (13) and (16) including $\boldsymbol{\Psi}_c$ which reduces the dimension of covariance matrix from $D \times D$ to $J \times J$, because $\hat{\mathbf{X}}^{\tilde{s}}$ can be poorly conditioned (i.e., ill-conditioned matrix). Thus, to alleviate the problem, after some matrix manipulations which multiplies on both sides of equations by $\hat{\mathbf{X}}^{\tilde{s}}$ before the inverse operation in (13) and (16), we estimated $\hat{\boldsymbol{\mu}}_c^{\tilde{s}}$ in a roundabout way instead of $\hat{\mathbf{Q}}_c$ or $\hat{\mathbf{z}}_c$. In that case, jBIT-MAPLR family becomes similar to MAP$_\text{BIT-MAPLR}$ family.

6. Conclusion

This paper proposes a new joint MAPLR framework using the single prior with a variance function where the speaker adaptation with scalable rectangular TM and canonical model is performed by tight coupling between direct and indirect methods. As future work, the probabilistic model for the prior model in BITS will be examined to statistically describe the behaviors of two variations which can be regarded as random variables (or latent variables) under online adaptation scheme.

Table 2: Word accuracy (%) of (j)BIT-MAPLR family

Adaptation Framework	Number of adaptation words						
	1	5	10	20	30	40	50
MAP$_\text{SI}$	95.93	95.68	95.63	95.25	95.10	95.24	95.09
MAP$_{\text{BIT-MLLR}_T(37)}$	-	94.55	**98.25**	**98.72**	**98.79**	**98.76**	**98.81**
MAP$_{\text{BIT-MLLR}_P(37/37)}$	**97.24**	**97.94**	97.97	98.01	98.04	98.02	98.06
BIT-MAPLR(20)	95.01	98.03	98.41	98.54	98.54	98.53	98.53
BIT-MAPLR$_\nu$(20)	***97.54***	**98.17**	98.38	98.50	98.52	98.55	98.57
jBIT-MAPLR(20)	94.85	98.03	**98.42**	**98.55**	98.57	98.58	98.55
jBIT-MAPLR$_\nu$(20)	97.49	98.12	98.36	98.50	**98.61**	**98.60**	**98.64**
BIT-MAPLR(30)	92.80	98.12	98.44	98.61	98.63	98.63	98.64
BIT-MAPLR$_\nu$(30)	**97.50**	***98.31***	**98.54**	**98.68**	98.72	98.70	98.73
jBIT-MAPLR(30)	92.64	98.12	98.45	98.65	98.71	98.69	98.70
jBIT-MAPLR$_\nu$(30)	97.50	98.26	98.50	98.66	**98.76**	**98.74**	**98.78**
BIT-MAPLR(37)	86.55	98.09	98.59	98.77	98.76	98.76	98.75
BIT-MAPLR$_\nu$(37)	**97.35**	**98.23**	98.54	98.75	98.71	98.76	98.77
jBIT-MAPLR(37)	86.63	98.07	***98.64***	***98.80***	98.75	98.73	98.73
jBIT-MAPLR$_\nu$(37)	97.34	98.20	98.55	98.74	***98.79***	***98.80***	***98.79***

7. Acknowledgments

This work was supported by the Industrial Strategic technology development program, 10035252, Development of dialog-based spontaneous speech interface technology on mobile platform funded by the Ministry of Knowledge Economy(MKE), Korea.

8. References

[1] Chen, K.-T. and Wang, H.-M., "Eigenspace-based maximum a posteriori linear regression for rapid speaker adaptation", in Proc. ICASSP, 8(6):695-707, 2000.

[2] Leggetter, C. J. and Woodland, P. C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, 9(2):171-185, 1995.

[3] Gauvain, J.-L. and Lee, C.-H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Trans. on Signal Proc., 39(4):806-814, 1994.

[4] Kim, D. K. and Kim, N. S., "Bayesian speaker adaptation based on probabilistic principal component analysis", in Proc. ICSLP, 734-737, 2000.

[5] Chesta, C. and Siohan, O. and Lee, C.-H., "Maximum a posteriori linear regression for hidden Markov model adaptation", in Proc. Eurospeech, 1771-1774, 1998.

[6] Siohan, O. and Chesta, C. and Lee, C.-H., "Joint maximum a posteriori adaptation of transformation and HMM parameters", in Proc. ICASSP, 2945-2948, 2001.

[7] Song, H. J. and Kim, H. S., "Bilinear Model-Based Maximum Likelihood Linear Regression Speaker Adaptation Framework", IEEE Signal Processing Letters, 16(12):1063 - 1066, 2009.

[8] Song, H. J., Jeong, Y. and Kim, H. S., "Bilinear Transformation Space-based Maximum Likelihood Linear Regression Frameworks", in Proc. Interspeech, 2009.

[9] Tenenbaum, J. B. and Freeman, W. T., "Separating style and content with bilinear models", Neural Computation, 12(6):1247-1283, 2000.

[10] Dempster, A. P. and Laird, N. M. and Rubin, D. B., "Maximum likelihood from incomplete data via the EM algorithm", J. R. Statist. Soc., 39:1-38, 1977.