Probabilistic Bilinear Transformation Space-Based Joint Maximum A Posteriori Adaptation

Hwa Jeon Song, Yunkeun Lee, and Hyung Soon Kim

This letter proposes a more advanced joint maximum a posteriori (MAP) adaptation using a prior model based on a probabilistic scheme utilizing the bilinear transformation (BIT) concept. The proposed method not only has scalable parameters but is also based on a single prior distribution without the heuristic parameters of the previous joint BIT-MAP method. Experiment results, irrespective of the amount of adaptation data, show that the proposed method leads to a consistent improvement over the previous method.

Keywords: Speaker adaptation, bilinear model, MAP.

I. Introduction

Speech technology has recently been in the spotlight as an effective interface for such mobile services as voice-operated searches and dictation. These services continue to be customized to increase user satisfaction, a crucial component of which is speaker adaptation [1]-[7]. In fact, a speaker-adapted system must provide a consistent performance improvement in proportion to the amount of adaptation data. As an illustration of a simple but powerful expression satisfying this condition on a specific speaker, s, we may take

$$\boldsymbol{\mu}^s = \alpha\boldsymbol{\mu}^s_{SC} + \beta\boldsymbol{\mu}^s_{MLLR} + \gamma\boldsymbol{\mu}^s_{ML}, \quad (1)$$

where $\boldsymbol{\mu}^s$ denotes the adapted mean vector of speaker s; subscripts *SC*, *MLLR*, and *ML* denote the speaker clustering [2], maximum likelihood linear regression [3], and maximum likelihood estimates, respectively; and α, β, and γ, with the constraint $\alpha+\beta+\gamma=1$, are nonnegative and automatically adjusted according to the adaptation data size. That is, if the data size is very small, α becomes dominant. As the data size is increased, β is dominant, and as the data size continues to increase, γ asymptotically becomes 1. This property has been shown in many approaches [2], [4]-[6] and is mainly embodied under the maximum *a posteriori* (MAP) criterion [4]. However, each approach has proposed its own specific partial version of (1), but not a full version. A few such cases are described in III.

As a full realization of (1), we have already proposed a novel *joint* bilinear transformation space-based MAP linear regression (jBIT-MAPLR) framework [7], which is based on a bilinear transformation (BIT) using the bilinear model (BM) concept. However, although the prior model using the variance function [7] has a substantial effect on reaching the goal of (1), it contains some heuristic parameters. Thus, this letter proposes a more systematic jBIT-MAP framework using the prior model, based on a statistical approach.

II. BIT-Based Speaker Adaptation

First, assume that the D-dimensional mean vector of content c for speaker s is $\boldsymbol{\mu}^s_c = \mathbf{W}^s\boldsymbol{\xi}_c$, where $\mathbf{W}^s \in \mathbf{R}^{D\times(D+1)}$ is the affine transformation matrix (TM) for speaker s, $\boldsymbol{\xi}_c = [1 \quad \boldsymbol{\mu}^T_c]^T$ is the extended version of speaker independent (SI) mean vector $\boldsymbol{\mu}_c$, and $1\le c\le C$. Second, from the TMs of S speakers, two observation matrices for BIT are given as

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

where superscript VT denotes the vector transpose of any matrix; $\overline{\mathbf{W}}^s = \mathbf{W}^s - \mathbf{W}^0$ with the mean matrix $\mathbf{W}^0$. Then, the two matrices in (2) are decomposed as

Manuscript received Feb. 6, 2012; revised Apr. 25, 2012; accepted May 11, 2012.

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.

Hwa Jeon Song (phone: +82 42 860 5836, hwajeon@etri.re.kr) and Yunkeun Lee (yklee@etri.re.kr) are with the BigData Software Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Hyung Soon Kim (kimhs@pusan.ac.kr) is with the School of Electrical Engineering, Pusan National University, Busan, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.12.0212.0054

$$\overline{\mathbf{W}} = [\mathbf{M}^{VT}\mathbf{S}]^{VT}\mathbf{Q} = [[\mathbf{MQ}]^{VT}\mathbf{S}]^{VT}, \quad (3)$$

where $\mathbf{S}=[\mathbf{s}^1 \ldots \mathbf{s}^S]$ and $\mathbf{s}^s$ is a unique factor for speaker s; $\mathbf{Q}=[\mathbf{q}_0 \ldots \mathbf{q}_D]$ is the orthogonal basis of the eigenvector; and the sizes of $\mathbf{S}$, $\mathbf{Q}$, and $\mathbf{M}$ are $I\times S$, $J\times(D+1)$, and $(ID)\times J$, respectively, with $I \leq S$ and $J \leq (D+1)$. Finally, the reconstructed model for speaker s is given as

$$\boldsymbol{\mu}_c^s \cong \mathbf{X}^s\mathbf{z}_c + \boldsymbol{\mu}_c^0, \quad (4)$$

where $\boldsymbol{\mu}_c^0 = \mathbf{W}^0\boldsymbol{\xi}_c$, $\mathbf{X}^s = [\mathbf{M}^{VT}\mathbf{s}^s]^{VT}$, and $\mathbf{z}_c = \mathbf{Q}\boldsymbol{\xi}_c$; further details of this speaker adaptation scheme can be found in [7].

III. Probabilistic BIT

From (1), we know that the adapted model asymptotically converges to the speaker dependent (SD) model $\boldsymbol{\mu}_{ML}^s$ as the amount of adaptation data continues to increase. In this case, the system provides the best performance for the speaker. Moreover, this property is easily realized by the standard MAP approach [4] in which a canonical model is used instead of $\boldsymbol{\mu}_{SC}^s$ and β is always fixed as 0 in (1). Moreover, to alleviate the inherent drawback in MAP when the data size is small, it is very important to build a scheme for the utilization of prior knowledge. An approach [5] has been proposed in which the prior model is based on a probabilistic principal component analysis (PPCA) [8] and satisfies the conjugate pair constraint. That is, from (1), this approach can be described as the case in which $\boldsymbol{\mu}_{SC}^s$ is built by the PPCA and β is always zero.

However, the above approaches cannot meet the goal of (1) due to the limited property of the prior model. Actually, in [7], we presented a simple jBIT-MAPLR scheme for our goal. However, there is a weak point in [7] in that the prior model was defined using a heuristic method. Thus, as a way to obtain a systematic setup, we propose a novel probabilistic BIT (PBIT)-based model with flexible and scalable properties. To achieve this, assume first that the d-th column vector in the reconstructed TM of speaker s, $\mathbf{w}_d^s$, is defined as

$$\mathbf{w}_d^s = [\mathbf{Mq}_d]^{VT}\mathbf{s}^s + \mathbf{w}_d^0 + \boldsymbol{\varepsilon}, \quad (5)$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is a random noise vector independent of $\mathbf{s}^s$ and $\mathbf{q}_d$ with variance σ^2; $\mathbf{s}^s \sim N(\mathbf{0}, \mathbf{I}_I)$ and $\mathbf{q}_d \sim N(\mathbf{0}, \mathbf{I}_J)$; $N(\cdot)$ denotes the multivariate normal density; $\mathbf{I}$, $\mathbf{I}_I$, and $\mathbf{I}_J$ are $D\times D$, $I\times I$, and $J\times J$ identity matrices, respectively; and $\mathbf{w}_d^0$ is the d-th column vector in the mean matrix $\mathbf{W}^0$. Under these assumptions, the relations among $\mathbf{w}_d^s$, $\mathbf{s}^s$, and $\mathbf{q}_d$ are given by

$$p(\mathbf{w}_d^s \mid \mathbf{s}^s, \mathbf{q}_d) = N(\mathbf{w}_d^s; [\mathbf{Mq}_d]^{VT}\mathbf{s}^s + \mathbf{w}_d^0, \sigma^2\mathbf{I}), \quad (6)$$

where the estimated parameters are $\lambda = \{\mathbf{M}, \mathbf{S}, \mathbf{Q}, \mathbf{w}^0, \sigma^2\}$ and the optimal model parameter $\hat{\lambda}$ can be obtained as

$$\hat{\lambda} = \arg\max_{\lambda}[\log p(\overline{\mathbf{W}} \mid \lambda)], \quad (7)$$

where λ and $\hat{\lambda}$ denote the current and the reestimated parameter sets, respectively. However, there is no way to simultaneously estimate $\mathbf{s}^s$ and $\mathbf{q}_d$. Hence, we propose a PBIT building procedure based on the PPCA.

First, the PPCA used to build the probabilistic approach for the prior model based on BIT is briefly described. Let the observation sequence, the latent variable sequence, and a factor loading in PPCA be $\mathbf{T}=[\mathbf{t}_1 \ldots \mathbf{t}_k \ldots \mathbf{t}_K]$, $\mathbf{Z}=[\mathbf{z}_1 \ldots \mathbf{z}_k \ldots \mathbf{z}_K]$, and the basis matrix $\mathbf{L}$, respectively. In the PPCA, assume that the relation among $\mathbf{T}$, $\mathbf{Z}$, and $\mathbf{L}$ is defined as

$$\mathbf{t}_k = \mathbf{L}\mathbf{z}_k + \boldsymbol{\mu}^0 + \boldsymbol{\varepsilon}, \quad (8)$$

where $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \sigma_P^2\mathbf{I})$, $\mathbf{z}_k \sim \boldsymbol{N}(\mathbf{0}, \mathbf{I})$, and $\boldsymbol{\mu}^0$ denotes the mean vector of $\mathbf{t}_k$. Based on these assumptions, $\mathbf{t}_k$ is normally distributed and the conditional probability distribution of $\mathbf{t}_k$, given $\mathbf{z}_k$, can be derived as

$$p(\mathbf{t}_k \mid \mathbf{z}_k) = N(\mathbf{t}_k; \mathbf{L}\mathbf{z}_k + \boldsymbol{\mu}^0, \sigma_P^2\mathbf{I}), \quad (9)$$

where the set of parameters $\lambda = \{\mathbf{L}, \boldsymbol{\mu}^0, \sigma_P^2\}$ can be estimated by the expectation-maximization (EM) algorithm with auxiliary function $Q(\hat{\lambda} \mid \lambda) = E[\log p(\mathbf{T}, \mathbf{Z} \mid \hat{\lambda}) \mid \mathbf{T}, \lambda]$, and $E[\cdot]$ is the statistical expectation. For more details, refer to [8].

Now, by combining the BIT with two interchangeable variations with the PPCA scheme, we propose the PBIT building procedure as follows:

Step 0. Initialization step

(a) Initialize the elements of σ^2, $\mathbf{S}$, and $\mathbf{Q}$ randomly.

(b) Initialize $\mathbf{M}$ with $\mathbf{S}$, $\mathbf{Q}$, and $\overline{\mathbf{W}}$ from (3).

Step 1. Estimation of $\mathbf{S}$ based on the PPCA procedure

(a) Assignment of PPCA parameters with fixed Q.

$$\mathbf{L} \leftarrow [\mathbf{MQ}]^{VT}, \mathbf{T} \leftarrow \overline{\mathbf{W}}^{VT}\ \sigma_P^2 \leftarrow \sigma^2.$$

(b) After converging the PPCA procedure, we obtain some of the BM parameters as follows:

$$[\mathbf{MQ}]^{VT} \leftarrow \hat{\mathbf{L}}, \mathbf{S} \leftarrow \hat{\mathbf{L}}^T\mathbf{T}, \sigma^2 \leftarrow \hat{\sigma}_P^2, \mathbf{M} \leftarrow \hat{\mathbf{L}}^{VT}\mathbf{Q}^T.$$

Step 2. Estimation of $\mathbf{Q}$ based on the PPCA procedure

(a) Assignment of PPCA parameters with a fixed $\mathbf{S}$.

$$\mathbf{L} \leftarrow [\mathbf{M}^{VT}\mathbf{S}]^{VT}, \mathbf{T} \leftarrow \overline{\mathbf{W}}, \sigma_P^2 \leftarrow \sigma^2.$$

(b) After converging the PPCA procedure, we obtain some of the BM parameters as follows:

$$[\mathbf{M}^{VT}\mathbf{S}]^{VT} \leftarrow \hat{\mathbf{L}}, \mathbf{Q} \leftarrow \hat{\mathbf{L}}^T\mathbf{T}, \sigma^2 \leftarrow \hat{\sigma}_P^2, \mathbf{M}^{VT} \leftarrow \hat{\mathbf{L}}^{VT}\mathbf{S}^T.$$

Step 3. If the criterion is not satisfied, go to Step 1.

Herein, we assume that the mean vector $\mathbf{w}^0$ (or matrix $\mathbf{W}^0$) is fixed, as it is very hard to consistently estimate the value of

each element with respect to two variations for several reasons, including the difference in scale and the number between the row and column of the rectangular matrices in (2).

IV. Joint Probabilistic BIT-MAP Adaptation

Based on the previous procedure, we propose a new joint PBIT-MAP that meets our target of (1). Actually, the proposed method can be easily derived by substituting two prior models in jBIT-MAPLR in [7] with a single PBIT-based prior model.

However, in [7], the jBIT-MAPLR using a prior model with variance function (jBIT-MAPLR$_v$) has shown a better performance than jBIT-MAPLR. Hence, we also focus on the approach based on the jBIT-MAPLR$_v$ criterion in this letter. In fact, it fails to completely satisfy the requirement for a conjugate pair of the prior model. However, it is sufficient when the PBIT building procedure is directly applied to the SD models instead of TMs. The reason is briefly explained below.

First, $\boldsymbol{\mu}_c^s$ can be reconstructed by (4) and (5) as follows:

$$\boldsymbol{\mu}_c^s \cong [\mathbf{M}^{VT}\mathbf{s}^s]^{VT}\mathbf{Q}\boldsymbol{\xi}_c + \boldsymbol{\mu}_c^0 = [\mathbf{M}\mathbf{z}_c]^{VT}\mathbf{s}^s + \boldsymbol{\mu}_c^0 . \tag{10}$$

Second, we assume that the distribution of (10) based on (6) is defined as

$$p(\boldsymbol{\mu}_c^s \mid \mathbf{s}^s, \mathbf{z}_c) = N(\boldsymbol{\mu}_c^s; [\mathbf{M}\mathbf{z}_c]^{VT}\mathbf{s}^s + \boldsymbol{\mu}_c^0, \sigma^2\mathbf{I}) . \tag{11}$$

In fact, we can build the SD model in (10) directly from the training database (DB) and not by the linear transformation from the SI model. We can also estimate the parameters in (11) based on the PBIT building procedure. However, although this direct approach provides more detailed parameters than an indirect approach such as in (10), the number and dimensionality of the parameters in the latter are more compact and manageable than those in the former. Moreover, it is not easy to collect a sufficient DB to train the SD models. Hence, we can say that the assumption in (10) and (11) is not only reasonable but also more practical.

Finally, the auxiliary function with (11) as a prior model based on the EM algorithm is defined as

$$\begin{aligned} R(\hat{\lambda} \mid \lambda) = & \sum_{t=1}^{T}\sum_{c=1}^{C}\gamma_c(t)[-a(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_c^s)^T\boldsymbol{\Sigma}_c^{-1}(\mathbf{o}_t - \hat{\boldsymbol{\mu}}_c^s)] \\ & + \sum_{c=1}^{C} E[-b\left\|\hat{\boldsymbol{\mu}}_c^s - [\mathbf{M}\mathbf{z}_c]^{VT}\mathbf{s}^s - \boldsymbol{\mu}_c^0\right\|^2 \mid \lambda], \end{aligned} \tag{12}$$

where a=1/2, b=ap with p=1/σ^2, $\gamma_c(t)$ is the *a posteriori* probability of being in content c at time t given $\mathbf{O}$={$\mathbf{o}_t$, t=1,…, T}, and $\boldsymbol{\Sigma}_c$ is the covariance matrix of content c and is assumed to be diagonal. By substituting elements of (12) with $\hat{\boldsymbol{\mu}}_c^s = \hat{\mathbf{X}}^s\hat{\mathbf{z}}_c + \boldsymbol{\mu}_c^0$, which represents the adapted models of a new speaker s based on (4) and (10), and ignoring all terms independent of the estimated parameters $\hat{\mathbf{X}}^s$ and $\hat{\mathbf{z}}_c$, (12) can be simply rearranged as

$$\begin{aligned} R(\hat{\lambda} \mid \lambda) = & -a\sum_{c=1}^{C}(\hat{\mathbf{z}}_c^T\hat{\mathbf{X}}^{sT}\mathbf{V}_c\hat{\mathbf{X}}^s\hat{\mathbf{z}}_c - \mathbf{k}_c^T\hat{\mathbf{X}}^{sT}\hat{\mathbf{z}}_c - \hat{\mathbf{z}}_c^T\hat{\mathbf{X}}^{sT}\mathbf{k}_c) \\ & - b\sum_{c=1}^{C}\hat{\mathbf{z}}_c^T\hat{\mathbf{X}}^{sT}\hat{\mathbf{X}}^s\hat{\mathbf{z}}_c + b\sum_{c=1}^{C}E\left[[\mathbf{M}\mathbf{z}_c]^{VT}\mathbf{s}^s \mid \lambda\right]^T\hat{\mathbf{X}}^s\hat{\mathbf{z}}_c \\ & + b\sum_{c=1}^{C}\hat{\mathbf{z}}_c^T\hat{\mathbf{X}}^{sT}E\left[[\mathbf{M}\mathbf{z}_c]^{VT}\mathbf{s}^s \mid \lambda\right], \end{aligned} \tag{13}$$

where $\mathbf{V}_c = \sum_{t=1}^{T}\gamma_c(t)\boldsymbol{\Sigma}_c^{-1}$ and $\mathbf{k}_c = \sum_{t=1}^{T}\gamma_c(t)\boldsymbol{\Sigma}_c^{-1}(\mathbf{o}_t - \boldsymbol{\mu}_c^0)$.

Although there is no way to maximize the likelihood with respect to $\hat{\mathbf{X}}^s$ and $\hat{\mathbf{z}}_c$ simultaneously, the parameters can be separately estimated using the iterative MAP principle as in [7].

First, for estimating $\hat{\mathbf{X}}^s$, we fix the remaining parameters to be estimated. From $\partial R(\cdot)/\partial\hat{\mathbf{X}}^s = 0$, the estimation formula is obtained as follows:

$$\sum_{c=1}^{C}[\mathbf{V}_c\hat{\mathbf{X}}^s\mathbf{Z}_c + p\hat{\mathbf{X}}^s\mathbf{Z}_c] = \sum_{c=1}^{C}[\mathbf{k}_c\mathbf{z}_c^T + p\overline{\mathbf{X}}_P^s\mathbf{Z}_c], \tag{14}$$

where $\mathbf{Z}_c = \mathbf{z}_c\mathbf{z}_c^T$, $\overline{\mathbf{X}}_P^s = [\mathbf{M}^{VT}\overline{\mathbf{s}}^s]^{VT}$, and $\overline{\mathbf{s}}^s = E[\mathbf{s}^s \mid \lambda]$ $= (\sigma^2\mathbf{I}_J + \mathbf{E}_J^T\mathbf{E}_J)^{-1}\mathbf{E}_J^T(\mathbf{w}^s - \mathbf{w}^0)$ with $\mathbf{E}_J = [\mathbf{M}\mathbf{Q}]^{VT}$ from the PBIT building procedure. Here, we can observe that the weight is smoothed by the statistical expectation.

Second, after differentiating $R(\cdot)$ with respect to $\hat{\mathbf{z}}_c$ with a fixed $\hat{\mathbf{X}}^s$, $\hat{\mathbf{z}}_c$ is given as

$$\hat{\mathbf{z}}_c = [\boldsymbol{\Psi}_c + p\mathbf{E}]^{-1}[\hat{\mathbf{X}}^{sT}\mathbf{k}_c + p\hat{\mathbf{X}}^{sT}\mathbf{X}_P^s\overline{\mathbf{z}}_c], \tag{15}$$

where $\boldsymbol{\Psi}_c = \hat{\mathbf{X}}^{sT}\mathbf{V}_c\hat{\mathbf{X}}^s$, $\mathbf{E} = \hat{\mathbf{X}}^{sT}\hat{\mathbf{X}}^s$, $\mathbf{X}_P^s = [\mathbf{M}^{VT}\mathbf{s}^s]^{VT}$, and $\overline{\mathbf{z}}_c = E[\mathbf{z}_c \mid \lambda] = (\sigma^2\mathbf{I}_J + \mathbf{E})^{-1}\hat{\mathbf{X}}^{sT}(\boldsymbol{\mu}_c^s - \boldsymbol{\mu}_c^0)$.

Now we must verify whether $\hat{\mathbf{X}}^s$ and $\hat{\mathbf{z}}_c$ meet our goal of (1); however, we decided to leave this step out of this letter as it has already been discussed in detail [7].

V. Experiments and Results

We conduct the training, adaptation, and evaluation under the identical experimental setup described in [7]. By training a DB collected from 80 persons, an SI hidden Markov model (HMM) consisting of 3,380 tied-states with four Gaussian mixtures per state is constructed at the triphone level, where we use 36-dimensional feature vectors consisting of their deltas and double deltas and 12 mel-frequency cepstral coefficients. We then obtain 80 SD TMs using an MLLR adaptation on each speaker from the SI HMM, and each TM is normalized by subtracting the average matrix. Finally, **S**, **Q**, and **M** for the standard BIT are obtained using the symmetric BM building procedure, while the parameters for PBIT are estimated using

Table 1. Word accuracy of several methods (%).

Adaptation framework	Number of adaptation words						
	1	5	10	20	30	40	50
SC	97.33	97.98	98.04	98.07	98.08	98.09	98.07
MLLR	-	94.58	98.28	98.70	98.74	98.77	98.78
MAP	95.93	95.68	95.63	95.25	95.10	95.24	95.09
$BIT_P(20)$	97.34	97.78	97.80	97.85	97.85	97.83	97.82
$BIT_T(20)$	39.78	97.60	98.44	98.64	98.65	98.61	98.64
jBIT(20)	97.49	98.12	98.36	98.50	98.61	98.60	98.64
jPBIT(20)	97.80	98.30	98.51	98.65	98.66	98.63	96.66
$BIT_P(30)$	97.39	97.92	97.97	98.04	98.01	98.03	98.03
$BIT_T(30)$	-	96.72	98.45	98.67	98.73	98.71	98.72
jBIT(30)	97.50	98.26	98.50	98.66	98.76	98.74	98.78
jPBIT(30)	97.72	98.34	98.62	98.74	98.77	98.75	98.78

$BIT_P(J)$=BIT-MLLR by projection; $BIT_T(J)$=BIT-MLLR by transform; jBIT(J)=joint BIT-MAPLR$_v$; jPBIT(J)=proposed method; J=the number of bases

the proposed algorithm in this letter. Here, S=80, D=36, and C=13,520 (=3,380×4). With another 70 speakers, we use 1 to 50 words for the adaptation in the supervised mode and 400 words for the evaluation on each speaker. Here, the word accuracy of the baseline system is 96.05% [7]. In Table 1, as the amount of adaptation data continues to increase, the typical performances of the SC, MLLR, and MAP are shown.

We can observe that the jPBIT family leads to consistent improvement over the jBIT family, irrespective of J. In particular, jPBIT with J=30 leads to recognition performance that is nearly comparable to or better than that of the MLLR for more than 20 adaptation words. Moreover, it provides considerable improvement for a small data size (1 to 10 adaptation words), achieving the maximum relative error reduction of 20% compared to the best performance of BITs. That is, since the performance in the fast speaker adaptation depends heavily on the quality of the prior model, we can say that jPBIT has a more outstanding capability in the fast speaker adaptation than other methods. Moreover, the jPBIT estimate converges asymptotically to the ML estimate.

Next, we will discuss the three practical issues related to implementation. That is, normalization is required to improve the numerical instability. First, in Steps 1-(a) and 2-(a) of the PBIT building procedure, $\sigma_P^2 \leftarrow \sigma^2 / \delta_1(\lambda)$ instead of $\sigma_P^2 \leftarrow \sigma^2$ is used for normalization against the numerical instability in the iterative learning procedure because there is a large difference in scale between variables relating to $\mathbf{S}$ and $\mathbf{Q}$. Second, since there is also a difference in scale between (6) and (11), σ^2 in (14) and (15) should be normalized with any value of $\delta_2(\lambda)$ according to the kind of parameters to be estimated, where we determine $\delta_1(\lambda)$ and $\delta_2(\lambda)$ empirically. Finally, as in [6] and [7], we also observe numerical instability in the inverse operation of the term, including $\mathbf{\Psi}_c$, which is not necessarily invertible. Thus, we directly estimate $\hat{\boldsymbol{\mu}}_c^s = \hat{\mathbf{X}}^s \hat{\mathbf{z}}_c$ instead of $\hat{\mathbf{z}}_c$ by multiplying both sides on the left by $\hat{\mathbf{X}}^s$ before the inverse operation in (15) as $\mathbf{A}[\mathbf{V}_c + p\mathbf{I}]\hat{\boldsymbol{\mu}}_c^s = \mathbf{A}\mathbf{k}_c + p\mathbf{A}\mathbf{X}_P^s \overline{\mathbf{z}}_c$ with $\mathbf{A} = \hat{\mathbf{X}}^s \hat{\mathbf{X}}^{sT}$. That is, dimension reduction terms are excluded to improve the numerical stability. In such a case, the jPBIT family becomes similar to the MAP_{PBIT} family. However, developing a more analytic method estimating $\hat{\mathbf{X}}^s$ with a constraint such as $\hat{\mathbf{X}}^s \hat{\mathbf{X}}^{sT} = \mathbf{I}$ to obtain the stability is left as future work.

VI. Conclusion

This letter proposed a new joint BIT-MAP framework using the prior model, which statistically describes the behaviors of two variations as random variables. As a future work, we will adopt this framework not only for online adaptation schemes but also for feature space adaptation.

References

[1] Y. Cho and D. Yook, "Maximum Likelihood Training and Adaptation of Embedded Speech Recognizers for Mobile Environments," *ETRI J.*, vol. 32, no. 1, Feb. 2010, pp. 160-162.

[2] K.-T. Chen and H.-M. Wang, "Eigenspace-Based Maximum a posteriori Linear Regression for Rapid Speaker Adaptation," *Proc. ICASSP*, 2000, p. 317-320.

[3] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech Language*, vol. 9, no. 2, Apr. 1995, pp. 171-185.

[4] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, Apr. 1994, pp. 291-298.

[5] D.K. Kim and N.S. Kim, "Rapid Speaker Adaptation Using Probabilistic Principal Component Analysis," *IEEE Signal Process. Lett.*, vol. 8, no. 6, June 2001, pp. 180-183.

[6] O. Siohan, C. Chesta, and C.-H. Lee, "Joint Maximum a posteriori Adaptation of Transformation and HMM Parameters," *Proc. ICASSP*, 2001, pp. 2945-2948.

[7] H.J. Song, Y. Lee, and H.S. Kim, "Joint Bilinear Transformation Space Based Maximum a posteriori Linear Regression Adaptation Using Prior with Variance Function," *Proc. Interspeech*, 2011, pp. 2577-2580.

[8] M.E. Tipping and C.M. Bishop, "Mixtures of Probabilistic Principal Component Analyzers," *Neural Computation*, vol. 11, no. 2, 1999, pp. 443-482.