Bilinear Transformation Space-based Maximum Likelihood Linear Regression Frameworks

Hwa Jeon Song, Yongwon Jeong, Hyung Soon Kim
School of Electrical Engineering, Pusan National University, Geumjeong-gu Jangjeon 2-dong, Busan 609-735, Korea
{hwajeon, jeongy, kimhs}@pusan.ac.kr

Abstract
This paper proposes two types of bilinear transformation space-based speaker adaptation frameworks. In training session, transformation matrices for speakers are decomposed into the style factor for speakers’ characteristics and orthonormal basis of eigenvectors to control dimensionality of the canonical model by the singular value decomposition-based algorithm. In adaptation session, the style factor of a new speaker is estimated, depending on what kind of proposed framework is used. At the same time, the dimensionality of the canonical model can be reduced by the orthonormal basis from training. Moreover, both maximum likelihood linear regression (MLLR) and eigenspace-based MLLR are identified as special cases of our proposed methods. Experimental results show that the proposed methods are much more effective and versatile than other methods.

Index Terms: speech recognition, speaker adaptation, bilinear transformation, bilinear model, MLLR

1. Introduction
Many speaker adaptation techniques have been proposed to improve the performance of automatic speech recognizer (ASR) based on speaker-independent (SI) hidden Markov model (HMM). These speaker adaptation methods can be generally categorized into direct and indirect methods. Usually, indirect method based on global transform is more powerful than direct method based on local transform when the adaptation data size is relatively small. Maximum likelihood linear regression (MLLR) [1] and eigenvoice [2] are two typical realizations for indirect speaker adaptation. Here, the essential point in eigenvoice method is to adjust the number of eigenvoices by employing the dimension reduction technique (DRT), such as principal component analysis (PCA). Thus, eigenvoice is useful for very small amount of adaptation data while MLLR becomes effective as more data is available. Moreover, indirect framework like eigenvoice can be also applied on transformation matrices themselves instead of a set of speaker-dependent (SD) models. These examples are the eigenspace-based MLLR (ES-MLLR) [3] and cluster adaptive training (CAT) using transform-based clusters [4]. Especially, both transforms in MLLR and ES-MLLR are based on the same canonical model, such as SI model. However, there is a training procedure for the canonical basis vectors of transformation matrix in ES-MLLR but not in MLLR. Also, it is hard to directly apply the DRT to MLLR in the model space. If both MLLR and ES-MLLR can be merged into one single framework with the DRT which shares the canonical basis vectors as well as canonical model together in a specific model space, that is more effective and versatile than the above speaker adaptation frameworks. This paper proposes one approach based on decomposing a set of transformation matrices into two factors: style factor identified as speakers’ characteristics and the dimension control factor which can reduce the dimensionality of the canonical model. In fact, an approach, where an observation is decomposed into two factors, has been originally proposed based on bilinear model (BM) in [5] and that has been applied on speaker adaptation in [6]. However, the property of our method is completely different from the previous works [5, 6] since the decomposition procedure in our method is applied on the transformation matrices themselves rather than on the mean vectors of observations. Furthermore, it is revealed that both MLLR and ES-MLLR are special cases of our proposed methods.

In Section 2, bilinear model-based speaker adaptation approaches are briefly described. The proposed frameworks in the bilinear transformation space are described in Section 3. In Section 4, experiments for the evaluation and comparison to other methods are discussed and a conclusion is drawn in Section 5.

2. Bilinear Model-based Speaker Adaptation
BM [5] not only decomposes a set of observations into the style and content factors but also controls these two factors independently. Being linked with the speaker adaptation technique [6], two kinds of BM, namely symmetric and asymmetric, are briefly described in the following.

First of all, let’s assume that the training database is collected from S speakers and each individual SD HMM is composed of N states with K mixtures per state. Let \( \mu^c_{nk} \) be the D-dimensional mean vector of Gaussian mixture k in state n of speaker s. \( \mu^c_s \), however, is used instead of \( \mu^c_{nk} \) for notation simplicity, where \( 1 \leq c \leq C \) (\( = N \times K \)).

2.1. Symmetric bilinear model
Let \( \mu^e_{s,d} \) be the dth element of \( \mu^e_s \). Based on the symmetric BM (s-BM), \( \mu^e_{s,d} \) can be expressed as follows:

\[
\mu^e_{s,d} = \sum_{i=1}^{I} \sum_{j=1}^{J} m_{ij,s} x^i_j y^c_j = x^s \mathbf{M}_s y^c
\]  

(1)

where \( x^s \), whose element is \( x^i_j \), denotes the I-dimensional style parameter vector of speaker s; \( y^c \), whose element is \( y^c_j \), denotes the J-dimensional content parameter vector of content c; \( \mathbf{M}_s \in \mathbb{R}^{I \times J} \), whose element is \( m_{ij,s} \), is the dth matrix among D matrices which are independent of both \( x^s \) and \( y^c \). These matrices map the style space spanned by \( x^s \) and the content space spanned by \( y^c \) into the D-dimensional observation space. The
matrix-vector form can be also obtained as
\[ \mu_s^c = \sum_{i=1}^{I} x_i^c y_i^c, \]
where \( m_{i,j} \) is a D-dimensional vector composed of \( m_{i,j,d} \).

2.2. Asymmetric bilinear model

There are often cases where a new style or content cannot be described accurately as a linear combination of a few corresponding basis factors from training. In these cases, more accurate factor can be described by two asymmetric BMs (a-BMs) instead of a-BM. First, based on the style factor in the matrix-vector form of a-BM, one a-BM can be build by replacing the interaction term \( m_{i,j,d} \) with \( x_i^c y_j^d \). Here, \( m_{i,j,d} \) denotes that \( m_{i,j,d} \) varies with a style factor. Therefore, the matrix-vector form of a-BM can be modified by introducing the style-specific matrix as
\[ \mu_s^c = \sum_{i=1}^{J} x_i^c y_i^d = X^c Y^d \]
where \( X^c \) is a D-dimensional vector whose elements are \( x_i^c \) and \( X^c \in \mathbb{R}^{D \times J} \), whose element is \( x_i^c \), describes the style-specific linear mapping from the content space into the observation space. For the speaker adaptation of a new speaker with the style \( s \), \( X^s \) has to be estimated from adaptation data while \( y^d \) is fixed. Based on (2), the relation between \( X^c \) and \( y^d \) corresponds to the relation between transformation matrix and canonical model in MLLR. Thus MLLR estimation formula can be naturally expanded to the \( s \)-BM case.

This paper proposes two versatile speaker adaptation frameworks based on the asymmetric bilinear model for speaker adaptation of a new speaker with style \( s \). For the speaker adaptation of a new speaker with style \( s \), the proposed methods can be naturally expanded to the \( s \)-BM case. It will be briefly checked in Section 3.2.

Let \( \mu_c \) be the mean vector of mixture \( c \) in SI model with Gaussian mixtures. And let’s assume that the relation between \( \mu_c \) and \( \mu^c \) is an affine transformation as \( \mu^c = A^c \mu_c + b^c = W^c \xi_c \), where \( A^c \) and \( b^c \) and \( W^c \) denote a D-D transformation matrix, \( D \) dimensional bias offset vector and \( D \times (D+1) \) affine transformation matrix, respectively; \( \xi_c = [1 \ \mu^c]^T \) is the extended SI mean vector. For notation simplicity a global transformation is assumed per speaker in this paper. The extension to multiple transforms is trivial. Let \( W^s \) be also defined as follows:
\[ W^s = [b^c A^c] = [w_0^s \ w_1^s \cdots \ w_D^s] \]
where \( w_0^s = b^c \) and \( w_d^s \) denotes the \( dh \) column vector in \( A^c \).

3. Bilinear Transformation Space-based Speaker Adaptation Frameworks

This paper proposes two versatile speaker adaptation frameworks in the bilinear transformation space, based on BM concept. After some setup for decomposition process, two approaches for speaker adaptation are presented based on BMs with simpler training session than s-BM case. Of course, since s-BM is more general than a-BM and it subsumes two a-BMs, the proposed methods can be naturally expanded to the s-BM case. It will be briefly checked in Section 3.2.

3.1. Transformation-based Approach

First, the transformation-based method, called bilinear transformation space-based MLLR by transform (BIT-MLLR), is proposed. To build a set of the style-specific matrix-based bilinear transformation parameters like (2), \( s \) transformation matrices are firstly stacked as follows:
\[ \mathcal{W} = \begin{bmatrix} W_1^T & \cdots & W_s^T & \cdots & W_D^T \end{bmatrix}^T \]
where the size of \( \mathcal{W} \) is \((SD) \times (D+1)\). Then SVD is applied on the \( \mathcal{W} \) as
\[ \mathcal{W} = UDV^T = QT. \]

Finally, from \( UDV^T \) which is the result of SVD on a family of transformation matrices, the first \( J \) column vectors of \( UD \) are taken as \( T \) and the first \( J \) row vectors of \( V^T \) are taken as \( Q \). Here, \( T=\begin{bmatrix} T_1^T & \cdots & T_J^T \end{bmatrix}^T \in \mathbb{R}^{D\times J} \) and its submatrix \( T_s^c = \begin{bmatrix} T_{1,s}^c & \cdots & T_{J,s}^c \end{bmatrix} \) denotes the style-specific matrix of speaker \( s \); \( Q=\begin{bmatrix} q_0 & q_1 & \cdots & q_D \end{bmatrix} \) denotes the orthonormal basis of eigenvectors to control the dimensionality of the canonical model since the dimension of \( q_d \), \( J \), can be varied as \( 1 \leq J \leq D+1 \). Notice that \( Q \) is the invariant factor to be commonly used across all the speakers’ transformation matrices. Moreover, the transformation matrix of speaker \( s \) can be approximated by taking the first \( J \) eigenvectors \((J \leq D+1)\) as \( W^c \cong QT \). Therefore, the mean vector of content \( c \) of speaker \( s \) is \( \mu_s^c = W^c \xi_c \cong QT \xi_c \). Now, let’s discuss the relation between \( Q \) and the content factor in BM. In fact, \( Q \) does not explicitly express the content factor in speech, whose example is a phonetic unit in BM [6]. On the other hand, \( \bar{X}^c \) and \( W^c \xi_c \), which is the linear transformation of \( \xi_c \), through \( Q \), is more adequate expression for the content factor than \( Q \). From this, \( \bar{X}^c \) and \( \bar{Z}^c \) become alternative transformation matrix of speaker \( s \) and canonical model of content \( c \), respectively, like MLLR. Fortunately, since the dimensionality of \( \bar{Z}^c \) can be reduced due to the property of \( Q \), the number of parameters in \( \bar{X}^c \) becomes smaller than that of \( W^c \) in MLLR.

In adaptation session, the style-specific matrix of a new speaker with style \( \hat{s} \) has to be estimated with the fixed \( Q \). Then the transformation matrix \( \hat{W}^{\hat{s}} \) of the new speaker can be defined as \( \hat{W}^{\hat{s}} = T^\hat{s} Q \), where \( T^\hat{s} \) denotes the style-specific matrix for the new speaker. Thus the adapted model of the new speaker becomes
\[ \hat{\mu}^c = \hat{W}^{\hat{s}} \xi_c = T^\hat{s} Q \xi_c. \]

In this paper, let’s assume that the linear transformation is performed only on the mean vectors but not on the covariance matrix. Let \( O = \{q_1, \cdots, q_D\} \) be the adaptation data of the new speaker. Given observation vector sequence \( O \), the auxiliary function [7] (ignoring all terms independent of the estimated parameters), based on ML estimation (MLE), is defined
as

\[ Q(\lambda, \hat{\lambda}) = \sum_{t=1}^{T} \sum_{c=1}^{C} \gamma_t (\alpha_t - \hat{\mu}_s)^T \Sigma^{-1}_s (\alpha_t - \hat{\mu}_s). \]  

(8)

Here \( \lambda \) and \( \hat{\lambda} \) denote the current and the re-estimated model parameter sets, respectively; \( \gamma_t \) is the posterior probability of being in a content \( c \) at time \( t \) given \( \Omega; \Sigma_s \) is the covariance matrix in Gaussian of mixture \( c \). \( T^* \) can be found by substituting (7) into (8) and then setting \( \partial Q(\lambda, \hat{\lambda}) / \partial T^* = 0 \). After some manipulations, the following is obtained:

\[ \sum_{t=1}^{T} \sum_{c=1}^{C} \gamma_t \Sigma^{-1}_s \alpha_t Q^T_s = \sum_{t=1}^{T} \sum_{c=1}^{C} \gamma_t \Sigma^{-1}_s T^* \xi_s \xi_s^T Q^T_s. \]  

(9)

If the covariance matrix \( \Sigma_s \) is diagonal, then (9) can be solved for \( T^* \) using the method originally proposed in [1].

Moreover, note that \( W^T(sT^*Q) \) becomes the same with the standard MLLR case when \( J=D+1 \), since no dimension reduction or no information loss due to the transformation by \( Q \) from \( \xi_s \) is occurred. That is, because \( Q^T \) in both side terms of (9) is canceled due to \( Q^T \Sigma_s Q=I \), \( Q \) just operates as an invertible constant matrix. Here \( I \) is the \((D+1) \times (D+1) \) identity matrix. Therefore, BIT-MLLR\(_T\) encompasses MLLR as a special case.

### 3.2. Projection-based Approach

Now, the projection-based method, called BIT-MLLR by projection (BIT-MLLR\(_P\)), is proposed. This is based on (3). For this, the vector transposed version of \( W \) is defined as

\[ W^{VT} = \left[ W^{1VT} \cdots W^{VT} \cdots W^{SVT} \right] \]  

\[ (10) \]

where \([\cdot]^{VT}\) denotes the vector transpose operation of a matrix \([\cdot] \); the size of \( W^{VT} \) is \((D+1)D\times S; W^{VT}_s \) denotes a meta-vector built by concatenating the \( D+1 \) ordered column vectors of \( W^s \) in (4) as \( W^{VT}_s = [w_0^T \cdots w_d^T]^T \). Then \( W^{VT} \) is decomposed into two factors as

\[ W^{VT} = UDVT = CS \]  

\[ (11) \]

where the first \( I \) column vectors of \( UD \) and the first \( I \) row vectors of \( VT \) as in BIT-MLLR\(_T\) are assigned to \( C \) (as content-specific matrix) and \( S \) (as style factor), respectively; \( U \) becomes equivalent of the eigenvectors in ES-MLLR if \( W^{VT} \) is normalized by the reference model like ES-MLLR; \( C = [C_0^T \cdots C_d^T \cdots C_j^T] \in \mathbb{R}^{(D+1)D \times I} \) whose sub-matrix \( C_j \in \mathbb{R}^{D \times I} \) denotes the canonical basis vectors of the \( S \) column vectors whose element is \( w_j \); \( S = [s^T \cdots s^T \cdots s^T] \in \mathbb{R}^{I \times S} \) and \( s^T \) denotes the style factor of the speaker \( s \). Moreover, \( C \) in (11) can be further generalized by employing the dimension control factor \( Q \) in (6) into BIT-MLLR\(_P\) framework. After some manipulations, \( C \) is converted as

\[ E_J = \left[ C^{VT} Q^T Q \right]^{VT} = \left[ W^{VT} S^{VT} T^{VT} Q^T Q \right]^{VT}. \]  

(12)

Here, if \( J=D+1 \), \( E_J \) becomes the same with \( C \) since \( Q^T Q=I \). That is, the eigenvectors in ES-MLLR can be regarded as a set of basis vectors obtained by eliminating the only style factor from the observation matrix except for the scaling factor \( D \) in (11). Hence ES-MLLR is a special example of BIT-MLLR\(_P\). If \( J<D+1 \), a new set of eigenvector for BIT-MLLR\(_P\) is created, which is different from ES-MLLR since \( Q^T Q \neq I \). Therefore the transformation matrix of speaker \( s \), based on \( E_J \) instead of \( C \), can be approximated as \( W^{VT} E_J s^T \) by taking the first \( I \) eigenvectors in \( S \) and \( J \) eigenvectors in \( Q \).

For further discussion related with \( s \)-BM, let \( M = [W^{VT} S^{VT} T^{VT} Q^T]^{VT} \in \mathbb{R}^{D \times J} \) in (12). In fact, \( M \) means a simplified bilinear mapping matrix from the iterative \( s \)-BM building procedure in [5]. Thus, based on the property between \( s \)-BM and \( a \)-BMs, the relation between \( E_J \) and \( M \) is equivalent of that between a set of \( Y^e \) in (3) and a set of \( M_{ab} \) in (1). Moreover, \( T^* \) and \( T^\gamma \) in (6) can be rearranged by \( M \) and \( S \) in (11) as \( T^*=M^{VT} S^{VT} T^{VT} \) and \( T^\gamma=M^{VT} S^{VT} T^{VT} \). Here, \( M^{VT} \in \mathbb{R}^{I \times J \times D} \) is \([W^{VT}]^{VT} \mathbb{S}^{VT} \). For more details of these matrix operations, refer [5].

Now, let’s BIT-MLLR\(_P\) apply for the speaker adaptation. In this case, the adapted model of a new speaker can be obtained by estimating the style factor based on the fixed \( E_J \) as

\[ \hat{\mu}_s = [MQ_s]^{VT} s^T \]  

where, for convenience, \( E_J = [MQ_s]^{VT} \) whose size is \( D \times I \). If (13) is substituted into (8) and set \( \partial Q(\lambda, \hat{\lambda}) / \partial s^T = 0 \), the style factor \( s^T \) is found after some manipulations as

\[ s^T = \sum_{t=1}^{T} \sum_{c=1}^{C} \gamma_t E^T_J \Sigma^{-1}_s o_t \]  

\[ (14) \]

This is equivalent of MLED in eigenvoice family [2, 3] or the estimation formulae of weight vector of clusters in CAT [4].

### 4. Experiments and Results

The proposed methods are evaluated in vocabulary-independent isolated word recognition experiments. For training, Korean phonetically optimized words (POW) database provided by ETRI, Korea is used. The training set contains words spoken by 40 male speakers. The feature vector of 36 components, consisting of 12 Mel-frequency cepstral coefficients, their delta, and delta-delta coefficients, is derived every 10ms over 20ms Hanning windowed segments. The SI HMM is based on a set of triphones and each triphone is composed of three state continuous density HMM with one Gaussian mixture per state. And 4050 tied-state are built by tree-based clustering. After SI training, 40 linear transformation matrices are obtained using MLLR for each speaker. The average transformation matrix is calculated from these 40 matrices and subtracted from each transformation matrix to produce normalized transformation matrices like ES-MLLR. Then, two SVD-based decomposition methods are applied to obtain bilinear transformation parameters. Here, \( S=40 \) and \( D=36 \) in (5) and (10). Korean phonetically balanced words (PBW) database provided by SITEC, Korea whose data acquisition environment is different from POW database is used for adaptation and evaluation. 1 to 50 words from 10 male speakers are used for adaptation in supervised mode and the rest 400 words are used for evaluation. The word accuracy of SI model as the baseline is 95.78%.

The results for MLLR, ES-MLLR, bilinear model-based method (BIM) [6] and the proposed methods are shown in Fig. 1. In all experiments of this paper, \((J/J)\) in BIT-MLLR\(_F(J/J)\) and \((J)\) in BIT-MLLR\(_P(J)\) denote the number of eigenvectors in the corresponding schemes. And the global transformation and MLE is used in all methods. A set of eigenvectors in ES-MLLR was obtained by applying PCA on \( W^{VT} \) [3].
that ES-MLLR is a kind of BIT-MLLR$_P$. Moreover, since ES-MLLR is identical with BIT-MLLR$_P(1/37)$, the performance of BIT-MLLR$_P(1/37)$ is only shown. From the figure, for very small amount of adaptation data, BIT-MLLR$_P$ family including ES-MLLR show the best results due to the smallest number of parameters to be estimated. And BIT-MLLR$_P$ with small $J$ leads to the robust performance improvement regardless of the amount of adaptation data on the whole. It is also seen in the following table. As the size of adaptation data increases, MLLR, BIM and BIT-MLLR$_T$ provide significant performance improvement over BIT-MLLR$_P$. Note that MLLR is equivalent of BIT-MLLR$_T(37)$. Especially, BIT-MLLR$_T$ and BIM outperform MLLR with fewer number of model parameters due to its scalability. For instance, when $J=20$, the number of parameters in BIT-MLLR$_T$ ($=36 \times 20$) is decreased nearly 46% in comparison to that of standard MLLR ($=36 \times 37$). In fact, BIM-based scheme can be naturally expanded to BIM based on (3) as well as (2) like BIT-MLLR family. However, BIT-MLLR family is more compact than BIM family, which is equivalent of the relation between ES-MLLR and eigenvoice.

In Table 1, the performance of the proposed methods for various number of eigenvectors is shown. In the table, ‘-’ denotes the cases where transformation matrix in MLLR or style-specific matrix in BIT-MLLR$_T$ cannot be estimated for lack of adaptation data. However, BIT-MLLR$_T$ with the small $J$ (e.g., $J=1$) still works for limited adaptation data, unlike MLLR. Thus BIT-MLLR$_T$ can be used for fast speaker adaptation if $J$ is appropriately determined depending on data size. Furthermore, BIT-MLLR$_T$ with $J=D+1$ provides the same performance with standard MLLR in the above figure, as aforementioned. On the other hand, when $I$ is small (e.g., $I \leq 10$), BIT-MLLR$_P$ with small $J$ leads to some improvement over that with $J=D+1$. However, when $I$ is more than 20, BIT-MLLR$_P$ with $J=D+1$ leads to some improvement over that with small $J$, because the canonical basis vectors with $J=D+1$ are more accurate. In fact, the matrix $Q$ in BIT-MLLR family does not guarantee the discrimination improvement between canonical models after dimension reduction by $Q Q_T$. Also, there is no guarantee of discrimination improvement in the bilinear mapping matrix $M$. Therefore more discriminant $M$ and $Q$ have to be built for the additional performance improvement in the proposed methods, but it will be examined in our future work.

### Table 1: Word accuracy (%) of proposed methods with different number of eigenvectors

<table>
<thead>
<tr>
<th>Adaptation Framework</th>
<th>Number of adaptation words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>BIT-MLLR$_P(1/37)$</td>
<td>95.9</td>
</tr>
<tr>
<td>BIT-MLLR$_P(1/10)$</td>
<td>96.4</td>
</tr>
<tr>
<td>BIT-MLLR$_T(1)$</td>
<td>96.6</td>
</tr>
<tr>
<td>BIT-MLLR$_T(5/20)$</td>
<td>97.1</td>
</tr>
<tr>
<td>BIT-MLLR$_T(5)$</td>
<td>88.0</td>
</tr>
<tr>
<td>BIT-MLLR$_T(10/20)$</td>
<td>97.1</td>
</tr>
<tr>
<td>BIT-MLLR$_T(10)$</td>
<td>97.1</td>
</tr>
<tr>
<td>BIT-MLLR$_T(20)$</td>
<td>94.2</td>
</tr>
<tr>
<td>BIT-MLLR$_T(30)$</td>
<td>97.1</td>
</tr>
<tr>
<td>BIT-MLLR$_T(20/10)$</td>
<td>97.1</td>
</tr>
<tr>
<td>BIT-MLLR$_T(20)$</td>
<td>97.1</td>
</tr>
<tr>
<td>BIT-MLLR$_T(20/10)$</td>
<td>97.4</td>
</tr>
<tr>
<td>BIT-MLLR$_T(30)$</td>
<td>97.1</td>
</tr>
<tr>
<td>BIT-MLLR$_T(30/20)$</td>
<td>-</td>
</tr>
<tr>
<td>BIT-MLLR$_T(37)$</td>
<td>97.4</td>
</tr>
<tr>
<td>BIT-MLLR$_T(37/20)$</td>
<td>97.4</td>
</tr>
</tbody>
</table>

5. Conclusion

This paper proposes the new speaker adaptation schemes in bilinear transformation space, which perform the dimension reduction of the canonical model and the speaker adaptation simultaneously. Furthermore, both MLLR and ES-MLLR are identified as special cases of our proposed methods. Thus, the proposed methods are more versatile than MLLR or ES-MLLR. Future work will be to develop an effective method for building more discriminant $Q$ by introducing the discriminant analysis, such as linear discriminant analysis.

6. Acknowledgements

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Knowledge Economy of Korea.

7. References


