RAPID FEATURE SPACE MLLR SPEAKER ADAPTATION WITH BILINEAR MODELS

Shilei Zhang[1], Peder A. Olsen[2], Yong Qin[1]

[1]IBM Research - China, Beijing 100193
{slzhang, qinyong}@cn.ibm.com

[2]IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
pederao@us.ibm.com

ABSTRACT

In this paper, we propose a novel method for rapid feature space Maximum Likelihood Linear Regression (FMLLR) speaker adaptation based on bilinear models. When the amount of adaptation data is limited, the conventional FMLLR transforms can be easily over-trained and can even degrade the performance. In such cases, usually by introducing structural constraints on the FMLLR transformation, the original FMLLR adaptation method can be modified for rapid adaptation. The objective of our bilinear model is to introduce a prior knowledge analysis on the training speakers based on Singular Vector Decomposition (SVD), and to incorporate it in the decoding process. This can effectively reduce the number of free parameters of FMLLR transformation and achieve performance improvements even with limited adaptation data. The efficiency of the proposed algorithm is demonstrated with experiments on the Mandarin digital dataset and the Mandarin voice search dataset respectively.

Index Terms— Rapid speaker adaptation, bilinear models, FMLLR, SVD

1. INTRODUCTION

Mismatch between the training and testing conditions leads to loss of some performance based on well-trained models. Many state of the art adaptation methods can help compensate for speaker variability, channel variability and content variability. Generally speaking, the model-based adaptation algorithm can be divided into three categories [1], speaker clustering based method which includes eigen-space based methods, Bayesian based method such as maximum a posteriori adaptation, and transformed based methods, such as maximum likelihood linear regression adaptation. These model-based methods need to change the speaker-independent HMM parameters, which can be computationally expensive and requires storing significant amount of data for the adapted speaker-dependent models.

In this paper, we focus on feature space maximum likelihood linear regression (FMLLR), which applies a single linear transform to the features. This is preferable for online rapid adaptation application, where rapid adaptation refers to adaptation with a small amount of adaptation speech. When the amount of available adaptation data is limited to the decoding process, the conventional algorithms can be easily over-trained, and result in very small performance improvement. In such case, by introducing some structural constraints on the FMLLR transformation, the original FMLLR adaptation method can be modified for rapid adaptation. In [2], eigen FMLLR is proposed for online incremental speaker adaptation, where the adapted transformation matrix is constrained to be a linear combination of a small number of basis *super-vectors* obtained from a set of reference speakers. This reduces the number of free parameters to be estimated. But the eigen FMLLR method do not have a close form solution, so it needs to apply numerical methods. Feature space maximum a posteriori linear regression (FMAPLR) uses a Bayesian prior to smooth FMLLR. FMAPLR can achieve robustness to limited amount of adaptation data by incorporating a prior distribution that is learned on the training data, [3,4,5]. However, the performance and solution of FMAPLR depends very much on the choice of the prior density. In this work, we propose a novel method for FMLLR speaker adaptation under the bilinear model framework based on the Singular Value Decomposition (SVD) to effectively incorporate prior information and reduce the number of free parameters. In the model fitting process, the content basis vectors will be estimated based on an SVD from the standard FMLLR matrices of the training speakers. The adaptation process selects the dimensionality of the content basis vector and finds the best style matrix for a new speaker based on the expectation maximization (EM) algorithm.

The rest of the paper is organized as follows: In section 2 FMLLR is briefly introduced. Section 3 describes the concept and formulation of bilinear models for FMLLR. In section 4 experiments are presented and the results will be discussed. We will draw some conclusions in section 5.

2. FMLLR

Feature space Maximum Likelihood Linear Regression (FMLLR) has proved to be highly effective as a method for unsupervised adaptation to a new speaker or environment [6]. It requires only a single transform matrix and bias vector to be estimated, which is implemented through a linear feature space transform:

$$\hat{O}(\tau) = AO(\tau) + b = W\xi(\tau)\,. \qquad (1)$$

where $O(\tau)$ is the N -dimensional feature vector at time τ in the original feature space, and $\hat{O}(\tau)$ is the transformed feature. $W = [b \quad A]$ is an $Nx(N+1)$ matrix which maximizes the likelihood of the adaptation data. A is the NxN transformation matrix; b is the $Nx1$ bias term. $\xi(\tau) = [1 \quad O(\tau)^T]^T$ is the $(N+1)x1$ extended observation vector.

Assume the acoustic models uses diagonal covariances. The objective of FMLLR is to maximize the likelihood. The EM algorithm gives an auxiliary function that can be maximized with respect to W to yield an increase in the likelihood:

$$\theta(\Theta, \hat{\Theta}) = \beta \log(p_i^T w_i) - \frac{1}{2}\sum_{i=1}^{N}[w_i^T G^{(i)} w_i - 2w_i^T k^{(i)}]\,, \qquad (2)$$

where w_i is the transpose of the ith row of W. p_i is the transpose of the extended cofactor row vector $[0, c_{i1}, \cdots, c_{iN}]$ for the ith row and $c_{ij} = cof(A_{ij})$ where $j = 1, \cdots, N$ with N being the dimension of feature. The sufficient statistics for estimating the transformation are

$$G^{(i)} = \sum_{m=1}^{M} \frac{1}{\sigma_i^{(m)2}} \sum_{\tau=1}^{T} \gamma_m(\tau) \xi(\tau) \xi(\tau)^T \tag{3}$$

$$k^{(i)} = \sum_{m=1}^{M} \frac{1}{\sigma_i^{(m)2}} \mu_i^{(m)} \sum_{\tau=1}^{T} \gamma_m(\tau) \xi(\tau) \tag{4}$$

$$\beta = \sum_{m=1}^{M} \sum_{\tau=1}^{T} \gamma_m(\tau) \tag{5}$$

and

$$\gamma_m(\tau) = p(q_m(\tau) \mid \Theta, O_T), \tag{6}$$

where $q_m(\tau)$ is Gaussian component m at time τ. $\gamma_m(\tau)$ is the posterior probability of $q_m(\tau)$ given the current adaptation data $O_T = \{O(1), \cdots, O(T)\}$. M is the total number of components associated with corresponding hidden state.

Differentiating with respect to w_i^T yields:

$$\frac{\partial \theta(\Theta, \hat{\Theta})}{\partial w_i^T} = \beta \frac{p_i^T}{p_i^T w_i} - w_i^T G^{(i)} + k^{(i)T} = 0. \tag{7}$$

By using direct method over rows, we get iterative solution,

$$w_i^T = (\alpha p_i^T + k^{(i)T}) G^{(i)-1}, \tag{8}$$

where α is solved by an iterative procedure following the derivation in [6].

3. FMLLR USING BILINEAR MODELS

Bilinear model can factor out two independent variations for the underlying style and content factors of observations and express them into a model, which is useful in various applications [7, 8]. Next, we will briefly introduce the basic concept of bilinear models and investigate the FMLLR estimation with bilinear models.

3.1. Bilinear Models

There are two types in bilinear models: symmetric and asymmetric. Because it is difficult to divide the observation data into two independent feature spaces in many applications, such as speech recognition, asymmetric bilinear models are usually employed [8].

3.1.1. Symmetric bilinear model

Let y^{sc} denote a K-dimensional observation vector in style s and content class c. We assume y^{sc} is a bilinear function of style parameter a^s and content parameter b^c with dimensionalities I and J as follows:

$$y_k^{sc} = \sum_{i,j} w_{ijk} a_i^s b_j^c. \tag{9}$$

The w_{ijk} terms are independent of style and content and characterize the interaction of these two factors. Furthermore, equation (9) can be rewritten in vector form as

$$y_k^{sc} = a^{sT} W_k b^c. \tag{10}$$

Here, $W_k \in R^{IxJ}$ with elements $\{w_{ijk}\}$ and there are totally K matrices describes a bilinear map to the K-dimensional observation vector.

3.1.2. Asymmetric bilinear model

Asymmetric bilinear model can change the interaction term ω_{ijd} according to a style and this is more flexible than the symmetric bilinear model. The above equation can be modified as follows to introduce style-dependent mapping matrices:

$$y_k^{sc} = \sum_{i,j} \omega_{ijk}^s a_i^s b_j^c. \tag{11}$$

Defining a style-specific term $a_{jk}^s = \sum_i \omega_{ijk}^s a_i^s$, then

$$y_k^{sc} = \sum_j a_{jk}^s b_j^c. \tag{12}$$

If we denote $A^s \in R^{KxJ}$ with elements $\{a_{jk}^s\}$, the above equation can be rewritten in vector form as

$$y^{sc} = A^s b^c \tag{13}$$

3.1.3. Model fitting for asymmetric model

Given a labeled training set of K-dimensional observations in S styles and C content classes, the observation matrix $Y \in R^{SKxC}$ can be stacked by the style-content pairs, where the pair can be one observation or mean of many observation samples, and then be constructed in matrix form $Y = AB$ as equation (13). Here, $A \in R^{SKxJ}$ is a stacked style parameter matrix and $B \in R^{JxC}$ is stacked by content parameters, respectively:

$$A^T = [A^1, \cdots, A^S] \quad B = [b^1, \cdots, b^C] \tag{14}$$

The goal of model fitting is maximum likelihood estimation of the style and content parameters. If the numbers of observations in each style and content class is equal in the training set, the asymmetric model can be fitted by a Singular Vector Decomposition (SVD) [8]. Then the observation matrix can be decomposed as $Y = USV^T$. We can then define the style parameter matrix A to be the first J rows of US and the content parameter matrix B to be the first J rows of V^T. For the extrapolation task, we find the style matrix that best explains the data for the new style class based on the Maximum likelihood criterion. More details about the concept of bilinear models can be found in [8].

3.2. Speaker adaptation using bilinear models

3.2.1. Bilinear Model building for FMLLR

For describing the FMLLR matrix using bilinear models, "style" can be defined as speaker standing for the variation across speakers and "content" can be defined as the columns of FMLLR matrix standing for the variation within the speaker. Let N be the dimension of feature vectors, then the standard FMLLR matrix for speaker s is an $Nx(N+1)$ matrix, W_0 is the empirical mean of FMLLR matrices of training speakers, and the observation matrix is arranged as a $SNx(N+1)$ matrix:

$$W_s = \begin{bmatrix} b_1 & a_{11} & \cdots & a_{1N} \\ b_2 & a_{21} & \cdots & a_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ b_N & a_{N1} & \cdots & a_{NN} \end{bmatrix} \in R^{NxN+1} \quad (15)$$

$$\bar{M}_A = \begin{bmatrix} W_1 - W_0 \\ \vdots \\ W_s - W_0 \\ \vdots \\ W_S - W_0 \end{bmatrix}, 1 \le s \le S, \quad W_0 = \sum_{i=1}^{S} W_i \quad (16)$$

Then the bilinear model for the observation matrix is computed based on the stacked FMLLR transforms from the training database composed of S speakers. To find the optimal style and content parameters, the observation matrix $\bar{M}_A$ can be decomposed and expressed in the asymmetric bilinear model as $\bar{M}_A = USV^T = AB$ by SVD, where S is diagonal matrix whose elements are singular values arranged in descending order. Then, A is defined as the first J columns of US and B is defined as the first J rows of V^T. The stacked style parameter is $A \in R^{(SN)xJ}$; $A^s \in R^{NxJ}$ denote the sth speaker matrix; and $B \in R^{Jx(N+1)}$ is the content parameters.

3.2.2. Adaptation process

The goal of adaptation process is to get the style factor for a new specific speaker t in iterative solution using content vector B learned during training based on maximum likelihood criterion. Under the bilinear model framework, the observation can be represented as:

$$\hat{O}(\tau) = W_t \xi(\tau) = (W_0 + A_t B)\xi(\tau) \quad (17)$$

Assume the diagonal covariance matrices are being considered, the objective of the maximum likelihood criterion is to maximum the following auxiliary function with respect A_t

$$\begin{aligned} &\theta(\Theta, \hat{\Theta}) = \beta \log(p_i^T w_{ti}) \\ &-1/2 \sum_{i=1}^{N} [(w_{0i}^T + A_{ti}^T B) G^{(i)} (w_{0i}^T + A_{ti}^T B)^T - 2(w_{0i}^T + A_{ti}^T B) k^{(i)} \end{aligned} \quad (18)$$

Where A_{ti}, w_{0i}, w_{ti} are the transpose of the ith row of the transform A_t, W_0, W_t, respectively. Statistical parameters $G^{(i)}$ and $k^{(i)}$ are same with equation (3), (4).

Then ignoring all terms independent of A_{ti}

$$\theta(\Theta, \hat{\Theta}) = \beta \log(p_i^T w_{ti}) - 1/2 \sum_{i=1}^{N} [A_{ti}^T \hat{G}^{(i)} A_{ti} - 2 A_{ti}^T \hat{k}^{(i)}] \quad (19)$$

where

$$\begin{cases} \hat{G}^{(i)} = (B G^{(i)} B^T) \\ \hat{k}^{(i)} = B k^{(i)} - B G^{(i)} w_{0i} \end{cases} \quad (20)$$

Differentiating with respect to A_{ti}^T yields

$$\frac{\partial \theta(\Theta, \hat{\Theta})}{\partial A_{ti}^T} = \beta \frac{p_i^T B^T}{p_i^T w_{ti}} - A_{ti}^T \hat{G}^{(i)} + \hat{k}^{(i)T} \quad (21)$$

The optimization can be solved by using direct method over rows. Assuming that the equation (21) is equating to zero for row i, then

$$\begin{aligned} &\beta \frac{p_i^T B^T}{p_i^T w_i} = A_{ti}^T \hat{G}^{(i)} - \hat{k}^{(i)T} \\ &p_i^T w_i \hat{k}^{(i)T} \hat{G}^{(i)-1} + \beta p_i^T B^T \hat{G}^{(i)-1} = p_i^T w_i A_{ti}^T \end{aligned} \quad (22)$$

Rearranging yields

$$\begin{aligned} A_{ti}^T &= \hat{k}^{(i)T} \hat{G}^{(i)-1} + \frac{\beta}{p_i^T w_i} p_i^T B^T \hat{G}^{(i)-1} \\ &= (\alpha p_i^T B^T + \hat{k}^{(i)T}) \hat{G}^{(i)-1} \end{aligned} \quad (23)$$

To find α, substituting this expression for A_{ti}^T in equation (22), and yields

$$\alpha^2 p_i^T B^T \hat{G}^{(i)-1} (p_i^T B^T)^T + \alpha (p_i^T B^T \hat{G}^{(i)-1} \hat{k}^{(i)} + p_i^T w_{0i}) - \beta = 0 \quad (24)$$

There will be two possible solutions in α. The value will be selected that maximizes auxiliary function. It is worth noting that the above formulas derivation is similar to the standard MLLR solution [6] except for the additional terms related to the prior information of content vector B and empirical mean W_0.

3.2.3. Bilinear Model Dimensionality Selection

The key advantage of bilinear model is to incorporate the prior information of training dataset into content basis vectors B fixed duration the adaptation and effectively reduce the number of free parameters from $W \in R^{Nx(N+1)}$ to $A_t \in R^{NxJ}$ by selecting J. The model dimensionality J is critical to bilinear model and can be chosen in various ways: a) by the amount of adaptation data; b) by the corresponding larger singular values; c) by the development set. In this work, we propose to pre-select J based on the objective function values of the corresponding adaptation data before the normal adaptation process as follows:

1) the initial value of J is $N+1$;
2) do adaptation based on bilinear model, iteration number=1;
3) compute the value of auxiliary function as equation (19);
4) $J = J - step$, go to step 2).

Finally, we can select the J with maximum objective function value. For bilinear models, in extreme case, when we have zero adaptation frame count, the FMLLR matrix will simply be W_0. When we get more and more data, the impact of the prior will become smaller along with more adaptation data. In other words, when we select $J = N+1$, the bilinear model is the same as FMLLR.

4. EXPERIMENTS

The proposed method was evaluated under two different scenarios: speaker-based adaptation on the connected digits dataset and utterance-based adaptation on the voice search dataset.

4.1. Connected Digits Experiments

The spoken connected digit database was collected in cars under different speed conditions, e.g., parked, low speed, high speed. The training set consists of 61 hours of digit data from 1,189 speakers. All above data were used for training acoustic models for Chinese digits recognition. Experiments are conducted on two testing corpora recorded in the same conditions as the training data.

The first corpus consisted of 3 data sets each containing about 580 utterances ranging in length from 2 to 5 digits, corresponding to three speed environments: parked car (*T0_var1*), low speed (*T1_var1*) and high speed (*T2_var1*) respectively; speech data with 6-15 digits utterances used in this second group consisted of 3 data sets each containing about 580 utterances, corresponding to three speed environments: parked car (*T0_var2*), low speed (*T1_var2*) and high speed (*T2_var2)* respectively.

Table 1. Performance comparison in sentence error rate (SER)

SER(%)	*eT2.var1*	*eT2.var2*	*eT1.var1*	*eT1.var2*	*eT0.var1*	*eT0.var2*
SI	12.7	31.1	12.1	28.6	11.3	23.5
+FMLLR	10.3	27.5	11.3	24.8	10.2	20.0
+Bilinear	10.1	27.0	11.8	24.3	9.7	19.9

For the Chinese digits recognition system, the 16 KHz input signal is coded using 13-dimensional PLP features with a 25ms window and 10ms frame-shift; 9 consecutive frames are spliced and projected to 40 dimensions using LDA. Head-Body-Tail model set [9] is employed, which consists of 33 phonetic units and is represented by 407 state Gaussian mixture models containing 12099 Gaussians totally under maximum likelihood (ML) training. The hypothesis output of baseline system are used to do FMLLR adaptation and the second path decoding based on speaker adapted feature is carried out. Comparison results for standard FMLLR and FMLLR with bilinear models adaptation are shown in *Table 1*. The data duration is 18 seconds per speaker in *var1* dataset, while it is 38 seconds in *var2* dataset. The results show that FMLLR with the bilinear model can achieve higher reduction in SER than standard FMLLR except for one subset *eT1.var1*. In this experiment, the model dimensionality J is selected automatically by using the method mentioned in section 3.2.3 and the step is 5. For the testing set *eT1.var1*, closer inspection of the automatic selection results reveals that the values J of objective function for the speakers with relatively large performance gap between bilinear models and standard FMLLR are very close. When we set the fixed $J = 31$ for all speakers, we can get the relatively optimal SER result of 11.2%. We can see the model dimensionality selection is very critical to the performance of bilinear models and need more work to find the optimal selection method.

4.2. Mandarin Voice Search Database

The Mandarin speech transcription system for a voice search application consists of speaker independent (SI) decoding, and speaker adapted (SA) decoding [10]. For the baseline system, the telephone input signal is coded using 13-dimensional PLP features with a 25ms window and 10ms frame-shift. In the front-end, the features are mean and variance normalized for each utterance, and then LDA, MLLT (maximum likelihood linear transform) are applied. Here three-state, left-to-right HMMs are used to represent 162 phones. The HMM states are context-dependent and clustered into equivalence classes by using decision trees. The distributions of 5K states are modeled by a pool of 150K Gaussian densities. The SA acoustic models share the same basic topology with the SI model. For speaker adaptation, two feature-space methods, VTLN and FMLLR are used in the baseline system.

Table 2. Performance comparison in character error rate (CER)

CER	*Voice Search database*
SI +SA w/ FMLLR	15.20%
SI+ SA w/ Bilinear model	13.75%

We use about 1800 hours 8 KHz data including 34447 speakers for HMM training and bilinear model training. There is about 1.3 hour of free style test-data including 714 utterances recorded by different speakers randomly. The length of each utterance is about 6 seconds on average. In such case, we can not confirm the utterances coming from the same speaker or those uttered in a roughly stationary environment, so the adaptation process need to be carried out utterance by utterance. As shown in *Table 2*, the CER based on FMLLR with bilinear models can achieve 9.5% relative reductions compared with that with standard FMLLR.

5. CONCLUSIONS

The prior information of the speaker variability from training dataset can reduce the amount of adaptation data necessary from the new speaker by constraining the parameter space, and help to rapidly adapt general acoustic characteristic to a new speaker. In this work, FMLLR estimation with bilinear model method for rapid speaker adaptation is proposed. The FMLLR matrix of each speaker is expressed as two independent spaces - speaker style and column content, and these two spaces are connected through bilinear mapping function in bilinear model, which can effectively incorporate prior information and reduce the number of free parameters. Bilinear model can be viewed as a generalization of FMLLR. As future work, we will investigate the class-dependent bilinear model adaption by training the different bilinear model based on class information in training dataset, such as gender, environment and so on. We also intend to investigate combination between bilinear modes with FMAPLR. Another aspect of our future work will efficiently control the speaker number of training dataset to further improve the robustness and performance by speaker clustering methods in training process.

6. REFERENCES

[1] B. K. Mak, T. Lai, I. W. Tsang and J. T. Kwok, "Maximum penalized likelihood kernel regression for fast adaptation", *IEEE Transactions on Audio, Speech & Language Processing*, Vol. 17(7), pp. 1372-1381, 2009.

[2] X. D. Cui, J. Xue and B. Zhou, "Improving online incremental speaker adaptation with eigen feature space MLLR", in *ASRU* , pp: 136-140, Merano, Dec., 2009.

[3] W. Chou, "Maximum a posteriori linear regression with elliptically symmetric matrix variate priors," in *EuroSpeech*, Vol. 1, pp. 1-4, Budapest, Hungary. Sep., 1999.

[4] X. Lei, J. Hamaker and X. D. He, "Robust feature space adaptation for telephony speech recognition", In *ICSLP*, pp: 1743-1746, Belgium, August, 2006.

[5] K. Visweswariah, V. Goel, and R. A. Gopinath, "Structuring linear transformations for adaptation using training time information," in *ICASSP*, pp 585-588, Florida, May 2002.

[6] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition", *Tech. Rep.*, Cambridge University Engineering Department, May 1997.

[7] H. J. Song, Y. Jeong, and H. S. Kim, "A new method for speaker adaptation using bilinear model", in *ICASSP*, pp.4365-4368, Taipei, April, 2009.

[8] J. Tenenbaum and W. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, pp. 1247-1283, 2000.

[9] S. L. Zhang, D. N. Jiang and Y. Qin, "Utterance verification using improved confidence measures based on alignment confusion rate in Chinese digits recognition," in *ICASSP*, pp.1309-1312, Taipei, April, 2009.

[10] S. M. Chu et al., "The 2009 IBM GALE Mandarin broadcast transcription system", in *ICASSP*, pp: 4374 – 4377, USA, March 2010.