MODEL DIMENSIONALITY SELECTION IN BILINEAR TRANSFORMATION FOR FEATURE SPACE MLLR RAPID SPEAKER ADAPTATION

Shilei Zhang, Yong Qin

IBM China Research Lab, Beijing 100094
{slzhang,qinyong}@cn.ibm.com

ABSTRACT

Bilinear models based feature space Maximum Likelihood Linear Regression (FMLLR) speaker adaptation have showed good performance especially when the amount of adaptation data is limited. However, the model dimensionality selection is very critical to the performance of bilinear models and need more work to find the optimal selection method. In this paper, we present an empirical study on this issue and suggest using a piecewise log-linear function to describe the relationship between the relatively optimal dimensionality parameter and the variant amount of data. This relationship can be used to efficiently select the bilinear model dimensionality in FMLLR speaker adaptation with the variant amount of data for each test speaker to improve recognition performance on the English voice control dataset.

Index Terms— Model dimensionality selection, bilinear models, FMLLR, rapid speaker adaptation

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.

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 speech application, most server-based speech recognition systems avoid model space adaptation [2]. In this paper, we focus on feature space maximum likelihood linear regression (FMLLR), it is sometimes known as Constrained MLLR (CMLLR), 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 relatively small amount of adaptation speech (often less than 30 sec.). 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, or even degrade the performance. In such case, by introducing some structural constraints on the FMLLR transformation, the original FMLLR adaptation method can be modified for rapid adaptation. Various methods have been proposed for rapid speaker adaptation. In [3], feature space maximum a posteriori linear regression (FMAPLR) uses a Bayesian prior to smooth FMLLR, which can achieve robustness to limited amount of adaptation data by incorporating a prior distribution that is learned on the training data. Because of robust performance, we use it as baseline of rapid adaptation here. In [4], we have proposed 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. Meanwhile, the model dimensionality is critical to bilinear model and selection method need to be investigated to achieve the optimal performance. In this paper, we investigate how the model dimensionality affects the performance of bilinear models in speaker adaptation experiments, and then propose a piecewise log-linear formula to define the parameter for each test speaker based on the experimental observations.

The rest of the paper is organized as follows: In section 2 FMLLR & FMAPLR as baseline method 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 & FMAPLR ESTIMATION

2.1. FMLLR

FMLLR has proved to be highly effective as a method for unsupervised adaptation to a new speaker or environment [5]. It requires only a single transform matrix and bias vector to be estimated, which is implemented through a linear feature space transform:

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

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 auxiliary function used for the estimation of transformation parameters with respect to W within EM framework given by:

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

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 as follows:

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

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

$$\beta = \sum_{m=1}^{M}\sum_{\tau=1}^{T} \gamma_m(\tau), \quad \gamma_m(\tau) = p(q_m(\tau) \mid \Theta, O_T) \quad (5)$$

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:

$$\frac{\partial\theta(\Theta,\hat{\Theta})}{\partial w_i^T} = \beta\frac{p_i^T}{p_i^T w_i} - w_i^T G^{(i)} + k^{(i)T} = 0 . \quad (6)$$

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

$$w_i^T = (\alpha p_i^T + k^{(i)T})G^{(i)-1}, \quad (7)$$

where α is solved by an iterative procedure following the derivation in [5].

2.2. FMAPLR

The basic idea of FMAPLR [3] is to apply the maximum a posteriori framework to maximize the following auxiliary Q-function with suitable prior distribution $p(W)$:

$$Q_{MAP} = Q_{ML} + \log p(W) \quad (8)$$

When the feature transformation matrix W is assumed to follow an elliptically symmetric matrix variate distribution as equation 9), we can estimate the FMAPLR transform in the same iterative way as in FMLLR.

$$p(W) \propto \exp[-\frac{1}{2}\sum_{i=1}^{n}(w_i - M_i)^T \upsilon_i^{-1}(w_i - M_i)] \quad (9)$$

where M_i is the location parameter and υ_i is the scale parameter for w_i . M_i and υ_i are called the hypermeters of the prior distribution. Compared with FMLLR, the FMAPLR only need add the extra smoothing value to the standard FMLLR statistical parameters G and K with the prior knowledge about the transform distribution as follows:

$$\hat{G}^{(i)} = G^{(i)} + \upsilon_i^{-1}; \quad \hat{k}^{(i)} = k^{(i)} + \upsilon_i^{-1}M_i \quad (10)$$

3. FMLLR USING BILINEAR MODELS

Generally speaking, in the model fitting process, the content basis vectors of bilinear model 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. Refer to [4] for the details.

3.1. 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 W_s 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 $\bar{M}_A$ is arranged as a $SNx(N+1)$ matrix as follows:

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

$$\bar{M}_A = \begin{bmatrix} W_1 - W_0 \\ \vdots \\ W_s - W_0 \\ \vdots \\ W_S - W_0 \end{bmatrix}, 1 \le s \le S; \quad W_0 = \sum_{i=1}^{S} W_i \quad (12)$$

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 style matrix; and $B \in R^{Jx(N+1)}$ is the content parameters.

3.2. 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:

$$\hat{O}(\tau) = W_t\xi(\tau) = (W_0 + A_t B)\xi(\tau) \quad (13)$$

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 , where ignoring all terms independent of A_{ti}

$$\theta(\Theta,\hat{\Theta}) = \beta\log(p_i^T w_{ti}) - 1/2\sum_{i=1}^{N}[A_{ti}^T \hat{G}^{(i)} A_{ti} - 2A_{ti}^T \hat{k}^{(i)}] \quad (14)$$

where $\hat{G}^{(i)} = (BG^{(i)}B^T); \quad \hat{k}^{(i)} = Bk^{(i)} - BG^{(i)}w_{0i}$ (15)

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).

Differentiating with respect to A_{ti}^T yields

$$\frac{\partial\theta(\Theta,\hat{\Theta})}{\partial A_{ti}^T} = \beta\frac{p_i^T B^T}{p_i^T w_{ti}} - A_{ti}^T \hat{G}^{(i)} + \hat{k}^{(i)T} \quad (16)$$

The optimization can be solved by using direct method over rows. Assuming that the equation (16) is equating to zero for row i, then

$$p_i^T w_{ti} \hat{k}^{(i)T} \hat{G}^{(i)-1} + \beta p_i^T B^T \hat{G}^{(i)-1} = p_i^T w_{ti} A_{ti}^T \quad (17)$$

Rearranging yields

$$A_{ti}^T = (\alpha p_i^T B^T + \hat{k}^{(i)T}) \hat{G}^{(i)-1} \quad (18)$$

To find α, substituting this expression for A_{ti}^T in equation (17), and yields

$$\alpha^2 p_i^T B^T \hat{G}^{(i)-1} (p_i^T B^T)^T + \alpha (p_i^T B^T \hat{G}^{(i)-1} \hat{k}^{(i)} + p_i^T w_{0i}) - \beta = 0 \quad (19)$$

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 [5] except for the additional terms related to the prior information of content vector B and empirical mean W_0.

3.3. 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 $A \in R^{NxJ}$ by selecting J. We can see the model dimensionality J is critical to bilinear model and need more deep work to investigate the selection method for rapid speaker adaptation application. In [4], we tried some work to pre-select J based on the objective function values of the corresponding adaptation data, but that method is not robust since small amount of adaptation data can not guarantee exact computation of objective function.

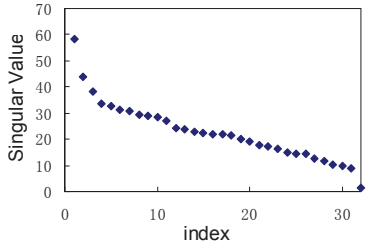

Fig.1. Singular values of observation matrix

As mentioned in section 3.1, S is a diagonal matrix with the singular values on the diagonal, which represent the relative importance of the corresponding content basis vectors. The descending order of s_i in matrix S implies that the *ith* content vector's contribution decreases as *i* approaches $N+1$. Intuitively but rather roughly speaking, a low rank approximation of FMLLR matrix can be obtained by discarding the least significant singular values and corresponding singular vectors. The singular values s_i extracted from observation matrix of training set in this paper are shown in Fig.1, and we also find that the plot is similar on other different task dataset. In the experiments, the values of bias term are much larger than those of transformation matrix in FMLLR estimation, so the largest singular value is much larger than others value and out of the range of Fig.1. The plot shows that the singular values decay slowly along the diagonal of the matrix (i.e. the singular values are relatively close), so we can not determine the J values by setting a threshold to select the top dominant singular values. On the other hand, generally speaking, the larger J value is, the closer approximate can be achieve to standard FMLLR. However, more number of free parameters need more amount of adaptation data. When the amount of available adaptation data is limited in the decoding process, the bilinear model with large J can be easily over-trained, and result in very small performance improvement. 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. Obviously, the setting of optimal J depends on the amount of data. We will discuss the dependence relationship of the model dimensionality J selection in depth in section 4.1.

4. EXPERIMENTS

The experiments were based on an English speech recognition system for voice control application including command & control, voice queries and short message dictation on mobile device. The acoustic model consists of 9k tied-states and 400k Gaussian components, trained on 4000 hours of data. The recognition features were 32-d vectors computed via an LDA+STC projection from 48-d MFCC features (the static cepstra plus the 1st, 2nd and 3rd order derivatives) with 10ms frame-shift. Model-space discriminative training was performed on the features. The language model used in the experiments was a general purpose trigram model. Bilinear model is trained on the FMLLR transformations of all the training speakers. Test set was composed of 40 speakers, referred to as *40 speakers set* here. To evaluate the fast adaptation performance with variant amount of adaptation data, separate adaptation data and test data sets were available for each test speaker. The adaptation data of each speaker was above 4 minutes long, recorded with same conditions of test data. True transcription by manual correction of decoding results generated by the decoder of baseline is employed on *40 speakers set*, treated as supervised adaptation scenario. Each speaker has about 2 minutes data, and it may have included various real-life background noises. The test set was used to study the effects of bilinear model dimensionality J parameter and evaluate the performance.

4.1. Model Dimensionality Selection

The setting of J depends on the amount of data. A larger amount of data requires a larger J to achieve the best performance, while a smaller amount of data needs a smaller J to reduce the number of free parameter and to improve the stability. In a speech system, the amount of adaptation data available for each speaker often varies, and using a fixed J values for all speakers will lead to suboptimal accuracy. Fig. 2 shows the performance comparison with the different J value for the variant amount of adaptation data including silence from 3 seconds to 1 minute on *40 speakers set*. In this experiment, we pick up data orderly at utterance level for each speaker, and end up with the data length larger than the setting amount of data (i.e. the adaptation data for each test speaker is just approximate to the setting data amount). It can be seen that when the adaptation data is relatively sufficient (i.e. >20 seconds), a larger J can lead to a lower WER, while for limited adaptation data (i.e. <20 seconds), a larger J will easily cause over-training, resulting in a very high WER. The smaller J reduces the number of free number effectively and guarantees the convergence for all amounts of data, especially for very limited data.

Fig. 3 shows the relatively optimal J obtained via manual tuning for each amount of adaptation data on test set (represented by blue line with square marks), where we tried different fixed J for all speakers and picked up the optimal J with best WER performance on each amount of data case. The plot shows that the optimal model dimensionality J of bilinear model highly depends on the amount of adaptation data, and the dependence tends to be piecewise log-linear. The J value decreases slowly when the amount of data is larger than 45seconds, while the plot declines relatively fast when the amount of data is smaller than 45 seconds. The mapping relationship between data amount and J can be described by the following piecewise log-linear functions:

$$\begin{cases} \ln(J) = 3.1457 + 0.043 * \ln(n) & n \geq 45s \\ \ln(J) = -1.2 + 0.8653 * \ln(n) & \text{otherwise} \end{cases} \quad (20)$$

where the linear parameters were learnt via linear regression. The correlation coefficient of $\ln(J)$ and $\ln(n)$ (n denotes the amount of data) was 0.99, confirming the log-linear dependence. The red line with dot marks in Fig. 3 shows the predicted J (rounding to integer) contour, which is rather close to the manually tuned one.

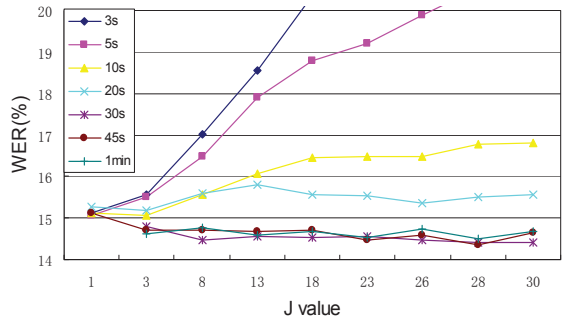

Fig. 2. Comparison of performance with different J value when the amount of adaptation data varies.

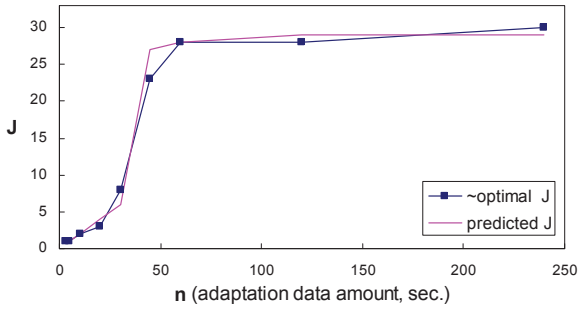

Fig. 3. Comparison of the best tuned J and the regressed J with the variant amount of adaptation data.

We also tried the similar experiment about model dimensionality selection on another dictation system, and get the same conclusion. The difference is the turning point. In the implementation, we can pre-compute the J values of some representative amount of data based on the prediction function, and then predict J by lookup table in estimation process. Then selection process will be very efficient and fast.

4.2. Experimental results

Tables 1 list the WERs of FMLLR, FMAPLR and bilinear model on *40 speakers set*. Standard FMLLR and FMAPLR are employed as baseline method, and two different J selection methods were compared: a) J was manually tuned to obtain the best accuracy specifically for each amount of data; b) J was predicted using (20). As mentioned in section 4.1, the exact amount of adaptation data for each speaker is slightly different on each amount of data case (especially when some utterances are long), therefore the bilinear model dimensionality J was predicted for each speakers in the bilinear experiments of *predicted J*, while the experiments results with *~optimal J* were based on the same dimensionality for all speakers as in Fig. 3. For both test sets, bilinear model with the J defined by (20) for each speaker (denoted as predicted J) significantly outperformed the FMLLR baseline for all amount of data and FMAPLR baseline for limited adaptation data (i.e. <=30s), and it performed similarly to bilinear model with J tuned manually. We also can see the conventional standard FMLLR transforms can be easily over-trained and can result in a very high WER, when there are relatively limited adaptation data (i.e. <=20s). On the other hand, the transformation estimated based on bilinear model can substantially achieve the robust performance even with very limited data. Especially, even when 3 seconds adaptation data were used for each speaker, bilinear model based on predicted J reduced the WER by relatively 3.4% compared with baseline result w/o adaptation for supervised adaptation scenario.

Table 1. Performance (WER) comparison with variant amount of adaptation data (baseline w/o adaptation: 15.66%)

40 speakers set	*1min.*	*30sec.*	*20sec.*	*10sec.*	*5sec.*	*3sec.*
standard FMLLR	14.56	14.64	16.01	17.43	22.62	25.51
FMAPLR	14.43	14.62	15.75	16.28	16.74	17.01
bilinear with ~optimal J	14.49	14.46	15.18	15.07	15.07	15.12
bilinear with predicted J	14.49	14.69	14.94	15.27	15.07	15.12

5. CONCLUSIONS

We present an empirical study that investigates the relationship between the dimensionality parameter and the variant amount of data based on the adaptation performance of bilinear models. The experimental results suggest that a piecewise log-linear function exists between the optimal model dimensionality and the amount of adaptation data, which could be used to predict dimensionality for a test speaker. With predicted J, bilinear models performance was close to that of bilinear model with J manually tuned, and better than standard FMLLR and FMAPLR, especially when limited adaptation data (<=30 sec) were available.

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] Neeraj Deshmukh, Puming Zhan, "Multi-Class Constrained Maximum Likelihood Linear Regression", Nuance Communications INC., U.S. Patent US 2009/0024390 A1, 2009.

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

[4] Shilei Zhang, Peder A. Olsen, Yong Qin, "Rapid feature space MLLR speaker adaptation with bilinear models", in *ICASSP*, pp: 4452-4455, Prague, Czech Republic, May 2011.

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