FAST SPEAKER ADAPTATION USING TRIPLE DIAGONAL AND SHARED BLOCK DIAGONAL TRANSFORM MATRICES

Guo-Hong Ding+ Bo Xu+ Juha Iso-Sipilä# and Yang Cao#*

High-Tech Innovation Center+, National Laboratory of Pattern Recognition*
Institute of Automation, Chinese Academy of Sciences, Beijing
{ghding, xubo}@hitic.ia.ac.cn
Nokia China R&D Center, Beijing#
{juha.iso-sipila, yang.1.cao}@nokia.com

ABSTRACT

This paper proposes two fast and effective adaptation algorithms, which are called SATD and SASBD respectively. The two algorithms are implemented in the MLLR frame and the transform matrices have constrained forms. SATD uses triple diagonal matrices to describe the mismatch between speakers and the acoustic model in the log-spectral domain and the matrices can be transformed into the cepstral domain to adjust the acoustic model. SASBD is different from the traditional block-diagonal MLLR and shares the three transformations of basic MFCC and dynamic features with one matrix. Moreover, both algorithms provide multiple choices for the biases. Experiments are extensively implemented and the results prove the advantages of SATD and SASBD over traditional MLLR.

1. INTRODUCTION

Many adaptation algorithms were proposed in the last decade. MLLR, MAP and extensions of the two algorithms were the mainstream. These days, fast speaker adaptation is more desirable, since it can provide the same or similar performance with fewer enrollment data.

The basic thought of fast speaker adaptation algorithms is how to utilize prior knowledge and characterize the variation among speakers with fewer parameters so that fewer enrollment data are needed to estimate the parameters. Here, "fewer enrollment data" is one of the aims of fast speaker adaptation and the other is the same or similar adaptation performance when enrollment data are enough. However, fewer parameters may describe the speaker variation partly or poorly. So, how to correctly and appropriately utilize prior knowledge is of importance for fast speaker adaptation.

Today, two tides of fast speaker adaptation techniques are speaker selection [4] and transform-based eigenvoice [2, 3]. The prior knowledge of the two algorithms lies in that specific speakers can be characterized by some parametric forms. In speaker selection technique, transform matrices or gaussian models are used to specify speakers and the distance of these parameters represent the distance of speakers. Utterances/ models of speakers close to the testing speaker are chosen to build the model. In transform-based eigenvoice, transform matrices are considered as the parametric form of speakers, the eigen transform matrices are obtained after analysis on the transform matrices of many speakers, and then the transform matrix of the testing speaker is the weighted sum of the eigen matrices.

Some other fast speaker adaptation algorithms use constrained parametric forms to compensate the mismatch between speakers and the acoustic model. For instance, VTLN uses warping factor to normalize the spectra of speakers to match the speaker-independent model [5]. [1] uses constrained transform matrices in the log-spectral domain to specify the variation among different speakers based on the prior knowledge of VTLN. [6] constrains the transform matrix with the combination of three rank-one vectors.

This paper proposes two fast and effective speaker adaptation algorithms. We call them SATD (Speaker Adaptation using Triple Diagonal transform matrices in the log-spectral domain) and SASBD (Speaker Adaptation using Shared Block Diagonal transform matrices) respectively. SATD inherits the triple diagonal form of the transform matrix in the log-spectral domain from the algorithm proposed in [1] and gives multiple choices for the biases. SASBD extends the traditional block-diagonal MLLR and shares the three transformations with one matrix. All the estimation formulas are developed and illustrated detailed in this paper

2. VTLN IN THE MLLR FRAMEWORK

In [1], VTLN is implemented in the MLLR framework. Since VTLN normalizes the feature by expanding or compressing the power spectra, the warped log-spectrum of a certain filter-bank component can be seen as a mapping of the unwarped. Since warping factor is usually close to unity, it can be assumed that the normalized log-spectrum of a certain component is a function of those of the neighboring three original components.

In [1], the function is linearly approximated according to the first-order Taylor series, so the normalized log-spectrum can be formulated as the linear weighted combination of three neighboring log-spectra and unity. Considering all components of the filter bank, we can formulate the linear functions as follows

$$L_\alpha = \theta^l L + b^l \tag{1}$$

where

This work was supported by National Key Fundamental Research Program of China (the 973 Program) under the grant G19980300504 and by National Natural Science Foundation of China under the grant 69835003.

$$\theta^l = \begin{bmatrix} \theta^l_{1,1} & \theta^l_{1,2} & 0 & \cdots & 0 \\ \theta^l_{2,1} & \theta^l_{2,2} & \theta^l_{2,3} & \ddots & 0 \\ 0 & \theta^l_{3,2} & \theta^l_{3,3} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \theta^l_{N,N} \end{bmatrix} \quad b^l = \begin{bmatrix} b^l_1 \\ b^l_2 \\ b^l_3 \\ \vdots \\ b^l_N \end{bmatrix} \tag{2}$$

$\theta^l_{i,j}$ and b^l_i are weighted coefficients, which can be taken fixed and represent the characteristics of the speaker.

Based on the property of DCT matrix, the transform matrix in the cepstral domain has the following relationship with the matrix in the log-spectral domain

$$\theta^c = D\theta^l D^T \tag{3}$$

$$b^c = Db^l \tag{4}$$

where D is $M \times N$ DCT matrix, which satisfies

$$DD^T = I_M$$

Although the triple diagonal transform matrix is derived from the approximation of VTLN, it can describe the variation between speakers more accurately since the constrained matrix contains much richer information than the warping factor.

3. THE PROPOSED ALGORITHMS AND THE ESTIMATION FORMULAS

SATD (Speaker Adaptation using Triple Diagonal transform matrices in the log-spectral domain) inherits the triple diagonal transform matrix in the log-spectral domain from the algorithm proposed in [1] and extend it with multiple reasonable forms of biases.

Moreover, motivated by SATD, as to the standard block-diagonal MLLR, we share the three transformations with one matrix, give multiple choices for the biases, and obtain a new speaker adaptation form, which is called SASBD (Speaker Adaptation using Shared Block Diagonal transform matrices).

3.1 SATD

The distinction between SATD and the algorithm proposed in [1] lies in the different forms of the biases.

As to transform-based speaker adaptation, we can make the assumption that dynamic features have similar or even the same affine matrix as basic MFCC since they are differences of the latter. If the biases are denoted differently for the three parts, the transform can be written as

$$\hat{\mu} = \theta\mu + b \tag{5}$$

where

$$\theta = \begin{bmatrix} \theta^c & & \\ & \theta^c & \\ & & \theta^c \end{bmatrix}, b = \begin{bmatrix} b^c_1 \\ b^c_2 \\ b^c_3 \end{bmatrix}$$

In the above formulas, θ^c is the transform matrix in the cepstral domain dependent on the triple diagonal matrix θ^l in the log-spectral domain (3).

Table 1 gives different choices for the biases. As table 1 shows, no biases are used in SATD-1 and it's the easiest form. Since dynamic features are obtained by difference, SATD-2 just considers the bias of basic MFCC. All biases are considered in SATD-3. From SATD-1 to SATD-3, they describe the variation more accurately, and need more parameters by degree.

SATD-1	$b^c_1 \equiv b^c_2 \equiv b^c_3 \equiv 0$
SATD-2	$b^c_1, b^c_2 \equiv b^c_3 \equiv 0$
SATD-3	b^c_1, b^c_2, b^c_3

Table 1: Several choices of biases for SATD

3.2 The estimation formulas of SATD

Construct the auxiliary function as follows

$$Q(\theta,\theta_o) = \sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_t(i)\left[K_i - \frac{1}{2}(y_t - \theta\mu_i - b)^T\Sigma_i^{-1}(y_t - \theta\mu_i - b)\right]$$
$$= \sum_{i=1}^{N}\frac{n_i}{2}\left[-(\overline{\mu}_i - \theta\mu_i - b)^T\Sigma_i^{-1}(\overline{\mu}_i - \theta\mu_i - b)\right] + K$$

where $\gamma_t(i)$ is the probability of output i at time t, K_i and K are independent of the transformation, and

$$n_i = \sum_{t=1}^{T}\gamma_t(i)$$

$$\overline{\mu}_i = \frac{1}{n_i}\sum_{t=1}^{T}\gamma_t(i)y_t$$

Since Σ_i is diagonal, $\overline{\mu}_i$, μ_i and Σ_i can be blocked

$$\Sigma_i = diag\{\Sigma^c_{i1}, \Sigma^c_{i2}, \Sigma^c_{i3}\} ,$$

$$\mu_i = \left[\mu^{cT}_{i1} \; \mu^{cT}_{i2} \; \mu^{cT}_{i3}\right]^T$$

$$\overline{\mu}_i = \left[\overline{\mu}^{cT}_{i1} \; \overline{\mu}^{cT}_{i2} \; \overline{\mu}^{cT}_{i3}\right]^T$$

Then if the independent constant is removed, the auxiliary function can be rewritten as follows.

$$Q(\theta,\theta_o) = -\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}\left[(\overline{\mu}^c_{ij} - \theta^c\mu^c_{ij} - b^c_j)^T\Sigma^{c-1}_{ij}(\overline{\mu}^c_{ij} - \theta^c\mu^c_{ij} - b^c_j)\right]$$
$$= -\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}\left[DD^T\overline{\mu}^c_{ij} - D\theta^l D^T\mu^c_{ij} - DD^T b^c_j\right]^T \cdot (\Sigma^c_{ij})^{-1}$$
$$\cdot\left[DD^T\overline{\mu}^c_{ij} - D\theta^l D^T\mu^c_{ij} - DD^T b^c_j\right]$$
$$= -\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}\left[(\overline{\mu}^l_{ij} - \theta^l\mu^l_{ij} - b^l_j)^T(\Sigma^l_{ij})^{-1}(\overline{\mu}^l_{ij} - \theta^l\mu^l_{ij} - b^l_j)\right]$$

where we denote

$$(\Sigma^l_{ij})^{-1} = D^T(\Sigma^c_{ij})^{-1}D , \tag{6}$$

$$\overline{\mu}^l_{ij} = D^T\overline{\mu}^c_{ij}$$

$$\mu^l_{ij} = D^T\mu^c_{ij}$$

If we define different forms of W^l and $\hat{\mu}^l_{ij}$ as table 2 displays,

SATD-1	$W^l = \theta^l, \hat{\mu}^l_{ij} = \mu^l_{ij}$
SATD-2	$W^l = \left[\theta^l \; b^l_1\right], \hat{\mu}^l_{i1} = \begin{bmatrix} \mu^l_{i1} \\ 1 \end{bmatrix}, \hat{\mu}^l_{i2} = \begin{bmatrix} \mu^l_{i2} \\ 0 \end{bmatrix}, \hat{\mu}^l_{i3} = \begin{bmatrix} \mu^l_{i3} \\ 0 \end{bmatrix}$
SATD-3	$W^l = \left[\theta^l \; b^l_1 \; b^l_2 \; b^l_3\right], \hat{\mu}^l_{i1} = \begin{bmatrix} \mu^l_{i1} \\ 1 \\ 0 \\ 0 \end{bmatrix}, \hat{\mu}^l_{i2} = \begin{bmatrix} \mu^l_{i2} \\ 0 \\ 1 \\ 0 \end{bmatrix}, \hat{\mu}^l_{i3} = \begin{bmatrix} \mu^l_{i3} \\ 0 \\ 0 \\ 1 \end{bmatrix}$

Table 2: The forms of W^l and $\hat{\mu}^l_{ij}$ according to the bias forms defined in table 1

then we can obtain

$$Q(\theta,\theta_o) = -\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}\left[(\overline{\mu}_{ij}^{l} - W^{l}\hat{\mu}_{ij}^{l})^{T}(\Sigma_{ij}^{l})^{-1}(\overline{\mu}_{ij}^{l} - W^{l}\hat{\mu}_{ij}^{l})\right]$$

Since W^l has special form as table 2 shows, some components of W^l are fixed to zeros. Differentiating $Q(\theta,\theta_o)$ with respect to W^l, removing those corresponding to the components fixed to zeros and letting the rest zeros yield

$$\sum_{i=1}^{N}\sum_{j=1}^{3} n_i\left[-(\Sigma_{ij}^{l})^{-1}W^{l}\hat{\mu}_{ij}^{l}\hat{\mu}_{ij}^{lT}) + (\Sigma_{ij}^{l})^{-1}\overline{\mu}_{ij}^{l}\hat{\mu}_{ij}^{lT}\right]_W = 0$$

where $[*]_W$ constrains matrix * to follow the form of W^l, that is, some fixed components of * are fixed to zero.

Since $(\Sigma_{ij}^{l})^{-1}$ isn't diagonal according to equation (6), we can't utilized the traditional method [8]. [1] provides a solution for the case.

3.3 SASBD

SASBD can be seen as an extension of the traditional block-diagonal MLLR. Following the same principle of SATD, SASBD shares the three transformations with one matrix as follows

$$\hat{\mu} = \theta\mu + b \tag{7}$$

where

$$\theta = \begin{bmatrix} \theta^c & & \\ & \theta^c & \\ & & \theta^c \end{bmatrix}, b = \begin{bmatrix} b_1^c \\ b_2^c \\ b_3^c \end{bmatrix};$$

We can notice that equation (7) has the same form as equation (5), but here, θ^c is full and all components are independent.

In the same way as SATD, SASBD has different choices for the biases defined in table 3.

SASBD -1	$b_1^c \equiv b_2^c \equiv b_3^c \equiv 0$
SASBD -2	$b_1^c, b_2^c \equiv b_3^c \equiv 0$
SASBD -3	b_1^c, b_2^c, b_3^c

Table 3: Several choices of biases for SASBD

3.4 The estimation formulas of SASBD

After the same proceedings as SATD, the auxiliary function can be written as

$$Q(\theta,\theta_o) = -\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}\left[(\overline{\mu}_{ij}^{c} - \theta^{c}\mu_{ij}^{c} + b_j^c)^{T}(\Sigma_{ij}^{c})^{-1}(\overline{\mu}_{ij}^{c} - \theta^{c}\mu_{ij}^{c} + b_j^c)\right]$$

Table 4 defines the forms of W^c and $\hat{\mu}_{ij}^{c}$ for SASBD-1, SASBD-2 and SASBD-3 respectively, and then we can obtain

$$Q(\theta,\theta_o) = -\sum_{i=1}^{N}\sum_{j=1}^{3}\frac{n_i}{2}\left[(\overline{\mu}_{ij}^{c} - W^{c}\hat{\mu}_{ij}^{c})^{T}(\Sigma_{ij}^{c})^{-1}(\overline{\mu}_{ij}^{c} - W^{c}\hat{\mu}_{ij}^{c})\right]$$

Differentiating $Q(\theta,\theta_o)$ with respect to W^c and letting it zero can yield

$$\sum_{i=1}^{N}\sum_{j=1}^{3} n_i[-(\Sigma_{ij}^{c})^{-1}W^{c}\hat{\mu}_{ij}^{c}\hat{\mu}_{ij}^{cT}) + (\Sigma_{ij}^{c})^{-1}\overline{\mu}_{ij}^{c}\hat{\mu}_{ij}^{cT}] = 0$$

In the above equation, $(\Sigma_{ij}^{c})^{-1}$ is diagonal, and we can estimate W^c by rows, as [8] proposed.

SASBD-1	$W^c = \theta^c, \hat{\mu}_{ij}^{c} = \mu_{ij}^{c}$
SASBD-2	$W^c = [\theta^c \; b_1^c], \hat{\mu}_{i1}^{c} = \begin{bmatrix} \mu_{i1}^{c} \\ 1 \end{bmatrix}, \hat{\mu}_{i2}^{c} = \begin{bmatrix} \mu_{i2}^{c} \\ 0 \end{bmatrix}, \hat{\mu}_{i3}^{l} = \begin{bmatrix} \mu_{i3}^{c} \\ 0 \end{bmatrix}$
SASBD-3	$W^c = [\theta^c \; b_1^c \; b_2^c \; b_3^c], \hat{\mu}_{i1}^{l} = \begin{bmatrix} \mu_{c1}^{c} \\ 1 \\ 0 \\ 0 \end{bmatrix}, \hat{\mu}_{i2}^{l} = \begin{bmatrix} \mu_{i2}^{c} \\ 0 \\ 1 \\ 0 \end{bmatrix}, \hat{\mu}_{i3}^{l} = \begin{bmatrix} \mu_{i3}^{c} \\ 0 \\ 0 \\ 1 \end{bmatrix}$

Table 4: The forms of W^c and $\hat{\mu}_{ij}^{c}$ according to the bias forms defined in table 3.

4. EXPERIMENTAL EVALUATION

The experiments are implemented on an isolated-word recognition system. Signals are sampled at 8kHz. 20ms speech frames are windowed and analyzed with FFT every 10ms. 20 mel filters are applied, and after logarithm and DCT, 12-d cepstra are obtained. The basic MFCC and its 1st-order and 2nd-order differences yield 36-d feature.

We record isolated words from 33 persons (16 females, 17 males), 108 words per speaker. To evaluate the proposed algorithm in the compensation for speaker variation, cross-gender experiments are implemented. Here, the acoustic model is built with utterances of male speakers and those of female speakers are used for test.

4.1 Experimental results of SATD

Since the number of the mel filters is 20, SATD-1, SATD-2 and SATD-3 have 58, 78 and 118 parameters to estimate respectively, according to the forms of θ^l and b^l.

Experimental results of SATD-1, SATD-2 and SATD-3 are given in figure 1. From the figure, we can conclude that the performances are very similar and that with only 2 enrollment isolated words, the adaptation performances achieve 95%, and then the performances increase very slowly.

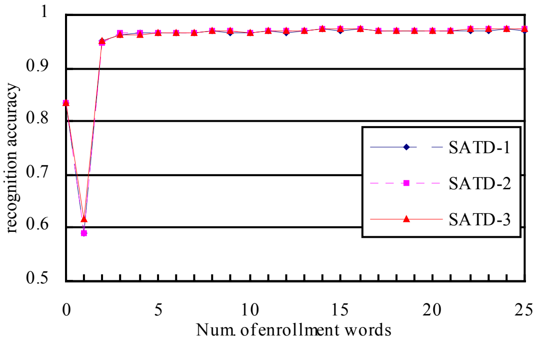

Figure 1: Recognition accuracy as a function of the number of the enrollment isolated words for SATD-1, SATD-2 and SATD-3

4.2 Experimental results of SASBD

SASBD-1, SASBD-2 and SASBD-3 have 144, 156 and 180 parameters to estimate respectively, according to table 4.

Figure 2 displays the experimental results of SASBD-1, SASBD-2 and SASBD-3. The figure shows that the performances of the three algorithms are also very similar. After about 7 enrollment isolated words, the performances can achieve 95%. Comparing figure 1 with figure 2, we can believe that

SASBD gives a little better performance than SATD with enough enrollment data.

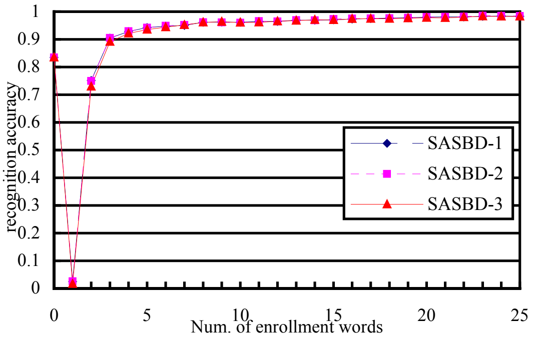

Figure 2: Recognition accuracy as a function of the number of enrollment isolated words for SASBD-1, SASBD -2 and SASBD-3.

4.3 Comparison of the proposed algorithms with the traditional MLLR adaptation algorithms

This section will give extensive experiments to compare the proposed algorithms with the traditional MLLR adaptation algorithms.

From figure 1 and figure 2, it's clear that the performances change little with respect to different cases of the biases. Here, we choose SATD-3 and SASBD-3 as the types for the two algorithms.

In the MLLR adaptation, transform matrices can be diagonal, block-diagonal or full. Usually, The diagonal matrix can lead to faster adaptation, while the block-diagonal or full transform matrices can bring about better performance. Here, the MLLR adaptation algorithms with the three kinds of transform matrices are chosen to experiment and evaluate.

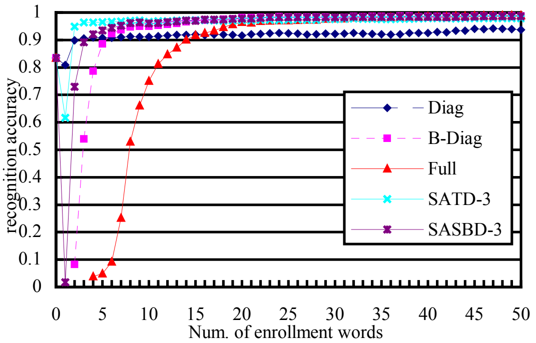

Figure 3: Recognition accuracy as a function of the number of enrollment isolated words for SATD-3, SASBD-3 and MLLR adaptation algorithms using diagonal (Diag), block-diagonal (B-Diag), and full transform matrices

The experimental results of the five algorithms are displayed in figure 3. There are some missing dots, such as the dots marking the recognition accuracy of the MLLR with the full transform matrix using 1 to 3 enrollment words and the reason lies in that too few enrollment data cause the estimation formulas singular.

From figure 3, we have following conclusions: 1) SATD-3 can provide the fastest adaptation. The performance approaches the best with only 2 or 3 enrollment isolated words and the best performance is very close to the best gained by all the algorithms. 2) MLLR with diagonal matrix only provides very limited adaptation performance, although it's also fast enough. 3) SASBD-3 is faster than the traditional MLLR with block-diagonal or full matrices. And it has the same performance as the latter two with enough enrollment data.

4.4 Discussion

The two algorithms proposed in the paper can be used in different conditions. When enrollment data are limited, SATD are useful, since it's much faster than SASBD and the MLLR with full or block-diagonal matrices. When enrollment data are sufficient, SASBD is preferred, since it's faster than the latter two and in the meanwhile, it has almost the same performance as MLLR with full or block-diagonal matrices when enough enrollment data are provided.

In the application of MLLR, regression class trees are used to generate multiple transformation matrices to match the specific speaker and the model more accurately. The two transformations proposed in the paper can be used to substitute the usual MLLR transformation.

Moreover, many techniques, such as speaker selection and transform-based eigenvoice, use transform matrices as the characteristics for specific speakers. The two transform forms proposed in the paper can be also used for these applications.

5. CONCLUSION

Two fast and effective adaptation algorithms are proposed in the paper. Both the two algorithms are transform-based and have constrained forms and fewer parameters. Experimental results prove the advantages of SATD and SASBD over the traditional MLLR.

6. REFERENCES

[1]. G.-H. Ding, Y.-F. Zhu, C. Li, B. Xu, Implementing Vocal Length Normalization in the MLLR Framework, In *Proc. ICSLP*, 2002.

[2]. K.-T. Chen, W.-W. Liau, H.-M. Wang, L.-S. Lee, Fast Speaker Adaptation Using Eigenspace-Based Maximum Likelihood Linear Regression, In *Proc. ICSLP*, 2000.

[3]. K.-T Chen, H.-M. Wang, Eigenspace-Based Maximum A Posteriori Linear Regression for Rapid Speaker Adaptation, In *Proc. ICASSP*, 2001.

[4]. C. Huang, T. Chen, E. Chang, Speaker Selection Training for Large Vocabulary Continuous Speech Recognition, In *Proc. ICASSP*, 2002.

[5]. L. Lee, R. Rose, A Frequency Warping Approach to Speaker Normalization, *IEEE Trans. Speech Audio Proc.*, 6: 49-60, 1998.

[6]. V. Goel, K. Visweswariah, R. Gopinath, Rapid Adaptation with Linear Combinations of Rank-One Matrices, In *Proc. ICASSP*, 2002.

[7]. J. McDonough, w. Byrne, etc., Speaker Normalization with All-Pass Transform, In *Proc. ICSLP*, 1998.

[8]. M. J. F. Gales and P. C. Woodland, Mean and Variance Adaptation within the MLLR Framework, *Computer Speech Language*, 10: 249-264, 1996.