COMPARISON AND STUDY OF SOME VARIANTS OF PARTIALLY TIED COVARIANCE MODELING

Peng Ding[1,2], Shuwu Zhang[1,2] and Bo Xu[1,2]

[1]High Technology Innovation Center; [2]National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, 100080
{*pding, swzhang, xubo*}@hitic.ia.ac.cn

ABSTRACT

In this paper, some practical implementation issues on *Partially Tied Covariance(PTC)* modeling are discussed. First, from the view of model complexity and computational load, a brief comparison is made for some variants of PTC. From the analysis, two representatives, STC[1] and Ortho-STC[2] are compared in details in the rest of the paper. Second, based on these variants, two techniques are studied. One technique is joint optimization of both transformation and HMM parameters, which will exploit the potential of PTC. And the other technique is model selection by hierarchical tree via *Bayesian Information Criterion*(BIC), which will decide the number and structure of transformation classes thus to assure the generalization capacity. Experiment results showed that STC always outperforms Ortho-STC due to the effect of parameter tying and by the application of above two techniques the system performance can be much improved.

1. INTRODUCTION

Covariance modeling for multidimensional speech data is a standard problem in acoustic modeling of the automatic speech recognition(ASR) systems. Typically, mixture of diagonal Gaussians are simply chosen to describe the acoustic features, with the underlying assumption that each element of the relatively high dimensional feature vector is independent. Since it is hard to find a single transform that de-correlates speech features of all states in a Hidden Markov Model(HMM) ASR system, the above assumption is always not true. The intra-frame correlations should be explicitly modeled, at least partly, to reinforce the system performance.

Efficient solutions should not only provide robust parameter estimation, but also have low memory and computational time overhead. The two extremes are diagonal-covariance model and the full-covariance one. Recently a number of tradeoffs among computational, storage and training data sparseness have been suggested and widely used in many state-of-the-art speech and speaker recognition systems[1-11]:

First category is the *Partially Tied Covariance(PTC)* modeling[1-8], a definition derived from[6]. They have in common that the full covariance(or its inverse) of each Gaussian component can be factored into the form ADA^T, where the state dependent untied parameter D is diagonal and the state independent parameter A is a linear transform, which is shared over a set of Gaussian components. They are different in the constraints putted on A and/or D.

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

The second one is *Factor Analyzed(FA)* covariance modeling [9, 10, 11], a kind of continuous-state linear Gaussian system modeling covariance structure of high dimensional static data using a small number of latent variables.

In addition, some variants can also be found: *covariance selection*[6] and *model selection*[7, 11]. The idea behind covariance selection is that by using a data-driven sparse structure on the covariance inferred from thresholding of the sample inverse covariance matrix, the unnecessary parameters of a system can be eliminated. As to the model selection, model structure[11], the assignment of the transformations[7] can be optimized.

Though there are some possible efficient algorithms for the likelihood computation involved in FA covariance models, the computational load is still much heavier than PTC models. So in this paper, we only concentrate on PTC and consider following problems:

1) Some variants of PTC are compared from the view of model complexity and computational load, based on the comparison we further choose the unconstrained STC[1] and *Common Principal Components*(CPC)[13] based orthonormal constrained STC(Ortho-STC)[2] for detailed comparison.
2) Two techniques are studied to optimize PTC, e.g., joint optimization of transformation and HMM parameters by multiple EM iterations[2] and model selection procedure on the number and structure of semi-tied transformations based on the *Bayesian Information Criterion*(BIC)[11].

The rest of the paper is organized as follows. In the following section, a brief review of some variants of PTC is described. In section 3, comparison and joint optimization of both transformation and HMM parameters for STC and Ortho-STC are made in depth, followed in Section 4, a new model selection procedure is introduced. Given all above, experiment results are presented in section 5, and we conclude in section 6.

2. COMPARISON OF SOME VARIANTS OF PTC

PTC modeling originates from various matrix decomposition methods for symmetric positive definite covariance matrix(or its inverse), e.g., $C=ADA^T$, where D is diagonal, matrix A can be of arbitrary form. According to the form of A, PTC can be categorized into two classes: *constrained* and *unconstrained*. For the former, some constraints are putted on A to remain mathematical restriction whereas for the latter, A is of any form.

2.1. Constrained PTC

Two variants are formerly used in speech recognition systems: *Eigen-decomposition*, named as State-Specific Rotation(SSR) in [3], and *Cholesky-decomposition*, named as Factored Sparse Inverse Covariance(FSIC) in [6]. In SSR, state dependent A is orthonormal and is composed by eigenvectors and D is made up

with real and positive eigenvalues. The feature space is rotated and all intra-frame correlations are removed. As in FSIC, A is a unit upper-triangular matrix that has ones along the diagonal and D is positive diagonal matrix. The matrix A describes intra-frame correlations, some elements can be forced to zeros by some covariance selection criterion. When A is tied across HMM states, the mathematical constraints can still be kept: SSR can be extended by Ortho-STC[2, 13], a CPC based simultaneous diagonalizing several weighted covariance algorithm, and FSIC can be extended by General EM algorithm as authors of [8] did.

2.2. Unconstrained PTC

There are no structural constraints putted on A. In addition to STC introduced in [1], some extensions have also emerged: 1) EMLLT[4] constrains the inverse covariance matrix of each Gaussian to be a linear combination of B rank one matrices. B is typically higher than the dimension of the feature d, if they are equal, the model moves back to global STC. 2) EMLLT model has been further extended to SPAM[5] by loosening rank one constraint for the basis subspace.

2.3. Comparison of Complexity and Computational Load

Covariance Models		Number Of Parameters	Computational Load
Diagonal		$M(2d + 1)$	$O(2Md)$
Full		$M(d(d + 1)/2 + d + 1)$	$O(M(d^2+d))$
Cons-Trained	Ortho-STC	$M(2d + 1) + Rd^2$	$O(2Md + Rd^2)$
	FSIC	$M(2d + 1)+Rd(d - 1)/2$	$O(2Md + Rd^2)$
Uncons-trained	STC	$M(2d + 1) + Rd^2$	$O(2Md + Rd^2)$
	EMLLT	$M(B + d + 1) + Bd$	$O(M(B + d) + Bd)$
	SPAM	$M(B + d + 1) + Bd (d + 1)/2$	$O(M(B + d) + Bd^2)$

Table 1: Comparison among some variants of PTC from the view of model complexity and computational load.

In table 1, let M be the total number of Gaussians of the system, which is typically on the order of 10,000, R denotes the number of transformation classes(also as semi-tied classes hereafter) and B, d are defined as those in section 2.2. From the table, it can be concluded that 1) all PTC models are compromises of reliable estimation of model parameters and computational cost between two extremes, e.g., diagonal and full covariance models; 2) both EMLLT and SPAM models are typically more computationally expensive than other PTC models.

We also believe that 1) there is some subtle relation between constrained and unconstrained PTC, in the following section, we take Ortho-STC and STC into detailed consideration of comparison; 2) with appropriate complexity control of the number of the model parameters, a balance leading to the best performance can be reached, which will be concerned in section 4.

3. JOINTLY OPTIMIZING OF TRANSFORMATION AND HMM PARAMETERS FOR ORTHO-STC AND STC

The unconstrained linear transformations may suffer from being unable to properly reflect the nature of the acoustic parameters[12], whereas when A is constrained to be orthonormal, it will lead to an explicit mathematical explanation as they are derived from the eigenvectors of the full covariance matrices. Moreover, with the orthonormal constraint, the *Jacobian* $|A|^{-1}$ equals to one, which means that the likelihood got in the original and the transformed spaces are directly comparable.

Because of the hidden nature of the state/mixture occupancy in HMM, an iterative two-step algorithm, expectation maximization(EM), can be used by maximizing the auxiliary function Q (1) with respect to HMM parameters and transformation: *(P, A)*.

Assuming $o(\tau)$ to be the d dimension observation at time τ, following auxiliary function must be optimized:

$$\mathbf{Q}(\hat{P},\hat{A}\,|\,P,A)=\sum_{m\in M^{(r)},\tau}\gamma_m(\tau)\times \{\log\hat{w}^m+\log\frac{|\hat{\mathbf{A}}^r|^2}{|\hat{\mathbf{\Sigma}}^m_{diag}|}-(\mathbf{o}(\tau)-\hat{\boldsymbol{\mu}}^m)^T\hat{\mathbf{A}}^{rT}\hat{\mathbf{\Sigma}}^{m^{-1}}_{diag}\hat{\mathbf{A}}^r(\mathbf{o}(\tau)-\hat{\boldsymbol{\mu}}^m)\} \quad (1)$$

where for the m^{th} ($m\in M^{(r)}$)Gaussian component of the r^{th} semi-tied class, the mixture weight, mean and diagonal covariance of the Gaussian component are denoted as w^m, $\boldsymbol{\mu}^m$ and $\mathbf{\Sigma}^m_{diag}$ respectively, and $\gamma_m(\tau)$ is state/mixture occupancy probability at time τ.

For both STC and Ortho-STC, sum of the mixture weights belonging to a state is constrained to be one. Especially for Ortho-STC, following constraint should also be satisfied for orthonormal property of A^r:

$$\mathbf{A}^r\cdot\mathbf{A}^{rT}=I$$

where I is d by d identity matrix.

According to EM algorithm, following procedure is summarized for the *joint optimization* of *(P, A)*:

1) Initialization: find an initial estimate of $P=\{w^m,\boldsymbol{\mu}^m,\mathbf{\Sigma}^m_{diag}\}$ and set A^r to be I;

2) E-step: compute the state/mixture occupancy probability at time τ : $\gamma_m(\tau)=p(m\,|\,o(\tau),P)$;

3) M-step: First using current estimation $\hat{\mathbf{A}}^r$ (or an identity matrix) to maximize (1) with respect to model parameters by taking partial derivative with $\hat{P}$, we obtain:

$$\hat{w}^m=\{\sum_{\tau=1}^T\gamma_m(\tau)\}/T \quad (2)$$

$$\hat{\boldsymbol{\mu}}^m=\{\sum_{\tau=1}^T\gamma_m(\tau)o(\tau)\}/\{\sum_{\tau=1}^T\gamma_m(\tau)\} \quad (3)$$

$$W^m=\{\sum_{\tau=1}^T\gamma_m(\tau)(o(\tau)-\hat{\boldsymbol{\mu}}^m)(o(\tau)-\hat{\boldsymbol{\mu}}^m)^T\}/\{\sum_{\tau=1}^T\gamma_m(\tau)\} \quad (4)$$

Second, estimate semi-tied transform $\hat{\mathbf{A}}^r$ using current set $\hat{P}$:

- **For STC**

A row by row optimization algorithm is introduced in [1], for a particular i^{th} row of $\hat{\mathbf{A}}^r$, $\hat{a}_i^{\,r}$:

$$\hat{\mathbf{\Sigma}}^m_{diag}=diag(\hat{\mathbf{A}}^rW^m\hat{\mathbf{A}}^{rT}) \quad (5)$$

$$\hat{a}_i^{\,r}=c_i(G^{ri})^{-1}\sqrt{\frac{\sum_{m\in M^{(r)},\tau}\gamma_m(\tau)}{c_i(G^{ri})^{-1}c_i^{\,T}}} \quad (6)$$

$$\mathbf{G}^{ri}=\sum_{m\in M^{(r)}}\frac{1}{(\hat{\sigma}^m_{diagi})^2}W^m\sum_\tau\gamma_m(\tau) \quad (7)$$

where c_i is i^{th} row vector of the cofactors of $\hat{\mathbf{A}}^r$, $\hat{\sigma}^m_{diagi}$ is element i of $\hat{\mathbf{\Sigma}}^m_{diag}$. Iteratively apply (5), (6) until convergence.

- **For Ortho-STC**

By applying the Lagrange multiplier for orthonormal constraint to (1)[2, 13], rows of $\hat{\mathbf{A}}^r$ are obtained by simultaneously solved from nonlinear equations united both (5) and (8):

$$\hat{a}_i^{rT}\{\sum_{m\in M^{(r)}}\frac{\hat{\sigma}^m_{diagi}-\hat{\sigma}^m_{diagj}}{\hat{\sigma}^m_{diagi}\hat{\sigma}^m_{diagj}}W^m\sum_\tau\gamma_m(\tau)\}\hat{a}_j^r=0\quad i\neq j \quad (8)$$

It can be seen that to attain a row vector and the corresponding diagonal elements satisfying orthonormal constraint, Nd+d(d-1)/2 equations are to be solved, assuming N is number of Gaussians in $M^{(r)}$. A very efficient algorithm, FG algorithm[13], can be adopted, which generalizes the *Jacobi* algorithm for simultaneously diagonalizing multiple weighted symmetric matrices.

4) Goto 2) until convergence.

It should be noted that in STC algorithm described in [1], only one EM iteration performed, followed by one or two Baum-Welch model parameters re-estimation, which differs from above procedure that both the transformation and HMM parameters are jointly re-estimated by multiple EM iterations.

4. MODEL SELECTION VIA BIC

It is known that models with too few parameters are always insufficient to model data complexity whereas too many leads to overtraining. Thus, the art of building a good acoustic model often lies in finding the right balance between the number of free parameters in the system and the amount of training data available. Parameter tying can be used to control model complexity and to improve the quality of estimates, hence to assure generalization capacity and to balance the computational loads as well.

In this paper, BIC, a likelihood criterion penalized by the model complexity, is used for model selection. Assuming we are given data set O and a set of candidates of desired parametric models Φ, then the optimal model should be selected by BIC for the given data[11]:

$$\hat{\Phi} = \arg\max_i (\log L(O,\Phi_i) - 0.5\lambda \times Ord(\Phi_i) \times \log(Nr(O))) \quad (9)$$

where $Ord(\Phi_i)$ is the number of parameters in model Φ_i, $Nr(O)$ is the number of training data. Typically, penalty weight $\lambda = 1$.

Here we apply BIC to determine the number of semi-tied transformations. The proposed method includes two steps: 1) apply STC/Ortho-STC for every HMM states; 2) build a bottom-up tree using a greedy search technique based on the least loss of the likelihood on the whole acoustic space[12]. The loss function can be herein defined as relative likelihood loss rate when certain state is transformed by other transformations. To build up the tree, starting from the whole set of HMM states with distinct transformation representing leaves, closest couple of states are first selected according to minimal loss criterion to define a new father node. A transform is then derived for the new node and the loss functions for all free nodes are renewed. This operation is repeated on the free nodes until the BIC value starts to drop.

Figure 1 illustrates how this procedure works. Here we only take STC for example, since the same trend can be seen for the case of Ortho-STC:

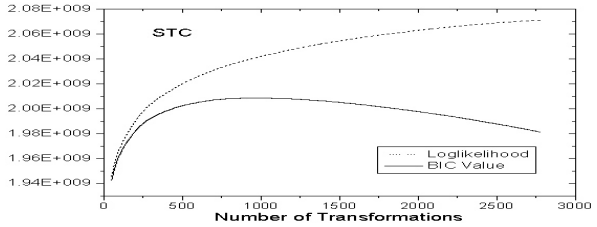

Figure 1: Scaled Likelihood and BIC value with the number of semi-tied transformations varying.

Clearly, with the increase of the model complexity, the likelihood always improves, whereas the BIC values first increases and then declines. Moreover, the BIC curve is even in a long range. In this paper, the optimal BIC value for STC is at 1,036 transformations and at 916 transformations for Ortho-STC.

At least three merits are enjoyed by the proposed technique: 1) model complexity and generalization can be well balanced via the imbedded BIC criterion; 2) a good starting point provided by the first step will benefit the following EM procedures; 3) components assignment can be found based on ML criterion.

5. EXPERIMENTS AND RESULTS

All experiments are carried out on our large vocabulary continuous mandarin speech recognition system[14]. Acoustic training database is composed of around 200,000 sentences of read, male/female balanced speech data recorded by about 500 speakers in quiet environments. The total number of phones(including silence as a separated phone) is 62, from which 2,776 distinct states were formed by a state clustered decision tree system. There were 16 Gaussian components used to model each state. A trigram language model is used in all the tests with a 40,000 words vocabulary. Other settings, including acoustic front-end, HMM topology, were the same as described in [14].

The testing set includes self-recorded, male/female balanced 1,200 utterances read by 20 speakers. The error rate is evaluated by character(CER) and is averaged over all 20 testing speakers. CER for the baseline system is 19.3%.

5.1. Evaluation of the Joint Optimization of Semi-tied Transformations and HMM Parameters

In this section, the objective is to evaluate the joint optimization and to compare its performance with respect to the standard one EM iteration optimization alone. The transformation classes are determined empirically by 1) *global* transformation; 2) individual transformation per *state*; 3) *phone* level transformation where all components of all states from the same phone share same class. In figure 2, STC systems with different transformation classes are used to test the performance of joint optimization.

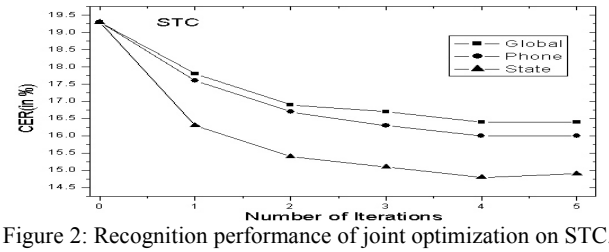

Figure 2: Recognition performance of joint optimization on STC.

For one thing, about further 8% relative error reduction typically can be seen after four iterations compared with the standard one-iteration STC. For example, a relative 8.8% vs. 17.1% error reduction can be obtained after one and four iterations respectively for phone level transform. The best performance in this test is a relative reduction of CER 23.3% obtained by state level transform after 4 iterations, indicating the effectiveness of joint optimization of both transformation and HMM parameters, which is well understandable by the nature of EM algorithms.

Another aspect is that with the increasing of the number of transforms, a consistent reduction of error rate can be shown. This seems to contradict the result of [7]: increasing the number of transforms showed no reduction in error rate. After a thorough comparison, we found that the training data set of [7] is composed of 36,493 sentences and by which a 6,399*12 Gaussians system is built, whereas 200,000 sentences to estimate 2,776*16 Gaussians in this paper. In heuristics, we attribute this disparity to the balance of model complexity and the amount of training data, e.g., how many transformations will best suit for a given task. This conjecture directly inspires the study of section 4.

5.2. Comparison of STC and Ortho-STC

The experiments presented in this section are aimed at comparing the performance and behavior of the unconstrained standard

STC with the constrained Ortho-STC under different model complexity, e.g., *global*, *phone* and *state* level. Best results by joint optimization formulation are used in this evaluation.

The results are illustrated in figure 3, which leads to the following comments: STC shows effectiveness over Ortho-STC under all three cases and the difference is reduced with the increase of the number of transformations. The trend explicitly shows the tradeoff between the mathematical rigor and the need of parameter tying. The more parameters tied, the more loss of performance brought by mathematic restrict will be seen. Analogous tendency can also be found in [2].

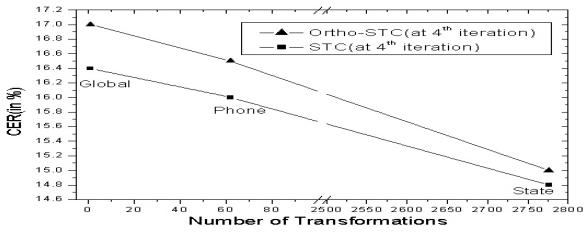

Figure 3: Comparison between STC and Ortho-STC with the number of semi-tied transformations varying.

5.3. Evaluation of Model Selection Strategy

In this section, performance of bottom-up clustering tree based model selection via BIC technique(named as *BICTree* hereafter) is presented. As a comparison, following routines are also implemented: Apply top-down cluster tree for components assignment by distance in acoustic space[10], followed by joint optimization(named as *RegTree* hereafter).

STC		Ortho-STC	
RegTree	*BICTree*	*RegTree*	*BICTree*
16.0 / 62	15.5 / 62	16.5 / 62	16.1 / 62
15.2 / 1036	14.6 / 1036	15.6 / 916	15.0 / 916
14.8 / 2776(*State*)		15.0 / 2776(*State*)	

Table 2: Performance comparison between routines of *BICTree* and *RegTree* respectively. The figures before slash in the table are CER in % and the others are the number of transformations.

Table 2 illustrated that the model selected by BIC are comparable or even better than the *state* systems with far more complexity, which will lead to more robust parameter estimation and less computational load. Another crucial aspect is that under the same model complexity, the system built by the proposed *BICTree* procedure outperformed the system built by above *RegTree* routines by about 0.4 - 0.6 in % absolute.

In figure 4 below, STC is taken for example to present the performance of the system built by *BICTree*. It is clear that the CER curve coincides with the BIC curve of figure 1 well. And also there is a long even range that the system kept in better performance, e.g., from about 500 to 2776 transformations herein.

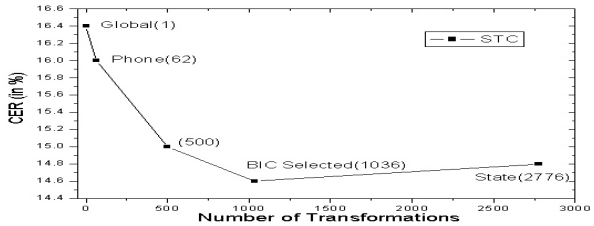

Figure 4: Performance of the STC system built by *BICTree* with the number of semi-tied transformations varying.

6. CONCLUSIONS

This paper has presented comparisons among some variants of PTC family from the view of model complexity and computational load. Based on the comparison, two representatives from both *unconstrained* and *constrained* categories respectively are compared in depth, e.g., STC and Ortho-STC.

Based on these variants, two techniques are also studied. The technique of joint optimization of both transformations and HMM parameters are introduced to exploit the potential of PTC, from which a further 8% relative error reduction has been reported compared with the standard one EM iteration scheme. To balance the amount of training data, model complexity thus to assure generalization, also balance computational load, an efficient model selection technique based on hierarchical tree via BIC is proposed to decide the number and structure of transformations, the performance of selected model are comparable or even better than the systems with far more complexity.

7. REFERENCES

[1] M. J. F. Gales, "Semi-Tied Covariance Matrices for Hidden Markov Models," *IEEE Trans. Speech and Audio Proc*. Vol. 7(3): 272-281, 1999.

[2] K. H. You and H. C. Wang, "Joint Estimation of Feature Transformation Parameters and Gaussian Mixture Model for Speaker Identification," *Speech Communication,* 3(1): 211-226, July 1999.

[3] A., Ljolie, "The Importance of Cepstral Parameter Correlations in Speech Recognition," *Computer Speech and Language*, Vol. 8: 223-232, 1994.

[4] P., Olsen and R. A. Gopinath. "Modeling Inverse Covariance Matrices by Basis Expansion," Submit to *IEEE Trans. Speech and Audio Proc*.

[5] S., Axelrod, R., A. Gopinath and P. Olsen, "Modeling with a Subspace Constraint on Inverse Covariance Matrices," In *Proc. ICSLP'02*, Denver, USA, 2002.

[6] Jeff. A. Bilmes, "Factored Sparse Inverse Covariance Matrices," In *Proc. ICASSP'00,* Istanbul, Turkey, 2000.

[7] M. J. F. Gales, "Factored Semi-Tied Covariance Matrices," In *Proc. NIPS 2000.*

[8] O., Cetin, H. J. Nock, K., Kirchhoff, J. Bilmes and M. Ostendorf, "The 2001 GMTK-Based SPINE ASR System," In *Proc. ICSLP'02*, Denver, USA, 2002.

[9] L., Saul, and M., Rahim, "Maximum Likelihood and Minimum Classification Error Factor Analysis for Automatic Speech Recognition," *IEEE Trans. Speech and Audio Proc*. Vol. 8(2): 115-125, 2000.

[10] P., Ding, Y., Liu and B., Xu, "Factor Analyzed Gaussian Mixture Models for Speaker Identification," In *Proc. ICSLP'02*, Denver, USA, 2002.

[11] S., S., Chen and R., A. Gopinath, "Model Selection in Acoustic Modeling," In *Proc. EuroSpeech,* 1999.

[12] C., Mokbel, "Online Adaptation of HMMs to Real-Life Conditions: A Unified Framework," *IEEE Trans. Speech and Audio Proc*. Vol. 9(4): 342-357, 2001.

[13] B., Flury, Common Principal Components and Related Multivariate Models, Wiley, New York, 1988.

[14] S., Gao, B., Xu, *et al*. "Update of Progress of Sinohear: Advanced Mandarin LVCSR System at NLPR," In *Proc. ICSLP'2000*, Vol.3, pp.798-801, Beijing, 2000.